Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference

de Waard, Dick; Toffetti, Antonella; Wiczorek, Rebecca; Sonderegger, Andreas; Röttger, Stefan; Bouchner, Petr; Franke, Thomas; Fairclough, Stephen; Noordzij, Matthijs; Brookhuis, Karel

Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference

Human Factors and User Needs in Transport, Control, and the Workplace

Edited by
Dick de Waard, Antonella Toffetti, Rebecca Wiczorek, Andreas Sonderegger, Stefan Röttger, Petr Bouchner, Thomas Franke, Stephen Fairclough, Matthijs Noordzij, and Karel Brookhuis
ISSN 2333-4959 (online)

Please refer to contributions as follows:
[Authors] (2017), [Title]. In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference (pp. pagenumbers). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Available as open access

Published by HFES

Contents

HMI AND USER EXPERIENCE

AVIATION

HEALTH

CONTROL ROOM PERFORMANCE AND OPTIMISATION

Investigation of driving behaviour reflecting drivers' risk anticipation for pedestrian collision risk assessment of right-turns at intersections

Hiroshi Yoshitake[1], Motoki Shino[1],
Hisashi Imanaga[2], & Nobuyuki Uchida[2]
[1]The University of Tokyo
[2]Japan Automobile Research Institute
Japan

Abstract

The objective of this study is to discover driving behaviour features that indicate pedestrian collision risk of right-turns at intersections for collision risk assessment. Right-turns at intersections is a typical accident scene of vehicle-pedestrian accidents in Japan, and if the collision risk of this scene becomes measurable, driver-assistance-systems will be able to prevent those accidents. To realise the objective, driving behaviour reflecting drivers' risk anticipation was focused on. Because collisions occur when drivers fail to anticipate risk correctly and driving behaviour is partially determined considering the drivers' anticipation, collision risk could be evaluated based on the driving behaviour features reflecting risk anticipation. To discover the driving behaviour features, first, near-miss incident data with high collision risk behaviour was analysed. Features of drivers' visual search and vehicle control were extracted as driving behaviour index candidates. Next, pedestrian collision risk scenes were experimentally reproduced and the relationship between the driving behaviour index candidates and collision risk were investigated. Evaluation indices based on drivers' visual search behaviour features showed significant correlation with pedestrian collision risk. From the results, it was clarified that pedestrian collision risk is evaluable with driver's visual search behaviour, and the visual search behaviour feature to realize the evaluation was identified.

Introduction

In Japan, among the fatal traffic accidents occurred in 2015, collisions with crossing pedestrians accounted for the most. Therefore, prevention of traffic accidents involving pedestrians is demanded to decrease the fatalities. If collision risk against pedestrians is evaluable in advance based on driving behaviour, driving-assistance-systems will be able to support the driver to select sufficient driving behaviour for collision avoidance and prevent collisions with crossing pedestrians. To realise such a system, a method to evaluate the collision risk against pedestrians based on driving behaviour is necessary.

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Drivers select their driving behaviour (e.g. adjust vehicle speed, adjust margins between objects) not to collide with static objects and other traffic participants based on the risk anticipation of the traffic environment (Van der Hulst et al., 1999). If driving behaviour selected to avoid collision based on driver's risk anticipation is identified, collision risk could be evaluated by comparing the present driving behaviour with the identified driving behaviour reflecting driver's risk anticipation. To anticipate the transition of traffic environment and set driving behaviour targets based on the anticipation, drivers use visual information obtained from their eyes. From this fact, not only driving operation and vehicle behaviour but driver's visual search behaviour is assumed to reflect the driver's risk anticipation and has the possibility to evaluate collision risk based on the behaviour. Therefore, visual search behaviour as well as driving operation and vehicle behaviour is focused on in this study.

Attention selection of driving is classified into four modes (Trick et al., 2004). Among the four modes, habit and deliberation are known as top-down selection, which are driven by goals and expectation (Engström et al., 2013). Habitual attention is often allocated to places where the main hazards exist based on the driver's experience, but in some situations if the attention allocation does not suit the actual situation it becomes critical. To avoid it, habitual visual attention is needed to be overridden by deliberate visual attention. To drive deliberate selection correctly and avoid the critical situation, sufficient risk anticipation of the situation is necessary. From this, it is suggested that the deliberate attention selection is related to risk anticipation. Therefore, when a driver is not anticipating the appearance of a pedestrian in a certain driving situation, it is assumed that the driver's visual attention will not be deliberate but habitual and the collision risk against the pedestrian will be high.

The objective of this study is to discover driving behaviour features reflecting risk anticipation of the driver which can evaluate pedestrian collision risk for future driver assistance systems, focusing on driver's visual search behaviour as well as driver's operation and vehicle behaviour. First, candidates of driving behaviour indices are extracted by analysing near-miss incident database containing high risk driving behaviour focusing on driving behaviour that reflects driver's risk anticipation. Next, the validity of the extracted driving behaviour indices is examined by a risk scene reproducing experiment using a real vehicle in a test course.

Target driving scene

The target driving scene in this study is right-turns at intersections with crossing pedestrians. Because vehicle-to-pedestrian collision during right-turns is a typical accident type in Japan and right-turn situation requires the driver to pay attention to many objects, this right-turn driving scene was selected as the target scene. In our previous research, environmental elements that affect driving behaviour of right-turns at signalized intersections with crossing pedestrians were clarified and 10 typical scene patterns were classified based on the clarified environmental elements (Shino et al., 2015). Among the classified typical scene patterns, the scene without

any preceding vehicle or oncoming vehicle (Pattern D), as shown in Figure 1, was set as the target driving scene in this study.

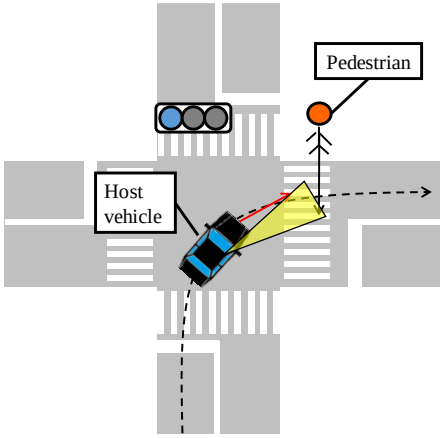


Figure 1. Bird's view of target driving scene (Pattern D).

Near-miss incident data analysis

The objective of this analysis is to extract driving behaviour indices reflecting driver's risk anticipation which has the possibility to indicate pedestrian collision risk in right-turns at intersections.

Method

To achieve the previously mentioned goal of this analysis, first, hypotheses of driving behaviour features reflecting driver's risk anticipation was formulated based on driving behaviours that drivers are expected to select when they have predicted the appearance of a crossing pedestrian at a crosswalk in a right-turn scene. Next, to examine whether the driving behaviour reflecting driver's risk anticipation have relation with pedestrian collision risk, the near-miss incident database collected by the Society of Automotive Engineers of Japan (JSAE) (Nagai, 2013) was used. This approach using a near-miss incident database approach is a valid approach because the composition of near-miss incidents and actual traffic accidents were similar and the near-miss events could be collected in larger volume. The database used contained over 80,000 near-miss events recorded by an acceleration triggered event recorder equipped on taxis running in 5 cities of Japan. The database stored recorded video images of the cameras on board, vehicle data (e.g. velocity, acceleration, GPS signals) and driver operation data (e.g. brake pedal on/off, turn signal). Because the visual attention of the driver is focused on in this study, data with two cameras (front camera and driver face camera) as shown in Figure 2 was used.

Figure 2. Sample camera image of JSAE near-miss incident database data (Left: Front camera, Right: Driver face camera).

Hypotheses of driving behaviour reflecting driver's risk anticipation

Driving behaviour reflecting driver's risk anticipation was extracted based on assumptions of driving behaviour that a driver will select to avoid collision in a situation where he/she anticipated the appearance of a pedestrian at an intersection. When a driver predicted the appearance of a crossing pedestrian in a right-turn situation, it is assumed that the driver will select their driving operation to avoid collision against the pedestrian with sufficient margin. As specific driving behaviour expected in the above situation, distributing visual attention to the ends of the crosswalk to find the pedestrian without any delay after the appearance and adjusting vehicle speed to maintain enough time to confirm the presence of the pedestrian can be listed. From the listed driving behaviour, distribution of visual attention to the surroundings and adjustment of vehicle velocity was focused on.

Result of database analysis

Figure 3 shows the time duration rate of each target that the drivers look during right-turns. The analysis period was from the timing when the vehicle crossed the centre line to the timing when the driver hit the brake against a pedestrian. The face direction of the driver in the above period was classified into 4 targets (traveling direction, oncoming direction, crosswalk and others) as shown in Figure 4. As a feature of the face direction during right-turns, a large time was turned towards the traveling direction. Looking long time towards the traveling direction will obstruct the driver from looking at other targets such as the crosswalk where pedestrians may appear and will lead to delay in crossing pedestrian perception. Figure 5 shows the frequency distribution of the velocity at the centre line. Incidents with velocity larger than 30 km/h accounted for more than the half. Crossing the centre line with high vehicle speed will shorten the time to check the appearance of pedestrians around the intersection and will lead to high pedestrian collision risk. From the analysis results, the face direction duration towards the traveling direction and the velocity at the centre line were extracted as driving behaviour indices.

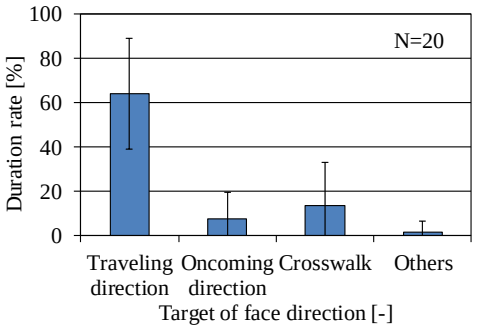


Figure 3. Time duration rate of face direction during right-turns.

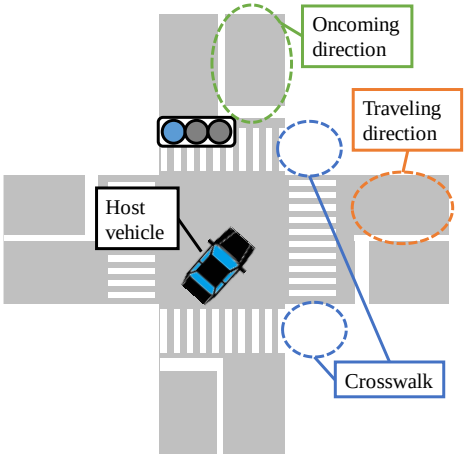


Figure 4. Targets of face direction during right-turns.

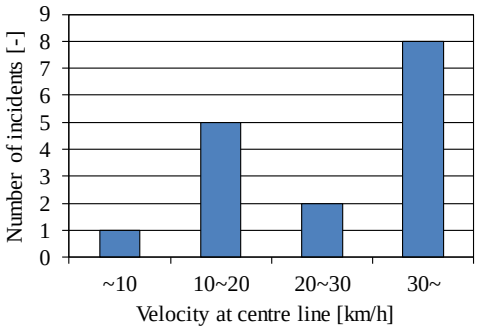


Figure 5. Frequency distribution of velocity at centre line.

Risk scene reproducing experiment

Objective

Driving behaviour indices reflecting driver's risk anticipation of pattern D was extracted based on analysis of the near-miss incident database. Although near-miss incident data classified as pattern D was analysed in this study, driving behaviour database contains data of various traffic environment conditions. Therefore, the objective of this reproducing experiment is to control environmental conditions by using a test course and examine the validity of the extracted indices.

Method and conditions

A reference measure TTC_{brake} was defined to evaluate the validity of the extracted driving behaviour indices reflecting the driver's risk anticipation. This TTC_{brake} is the time-to-collision value against a crossing pedestrian at the driver's braking manoeuvre timing calculated by the Equation 1 in the scene shown in Figure 6. D and θ in Equation 1 is defined as Equation 2 and 3. The correlation between the TTC_{brake} calculated based on a scene with a crossing pedestrian (risk scenario) driving data and the driving behaviour indices calculated based on a scene with no crossing pedestrian (non-risk scenario) driving data was evaluated.

$$TTC_{brake} = \frac{D-L}{V\cos\theta} \tag{1}$$

$$D = \sqrt{\left(X_{cp} - X_{hv}\right)^2 + \left(Y_{cp} - Y_{hv}\right)^2} \tag{2}$$

$$\theta = tan^{-1}\left(\frac{X_{cp}-X_{hv}}{Y_{cp}-Y_{hv}}\right) - \psi \tag{3}$$

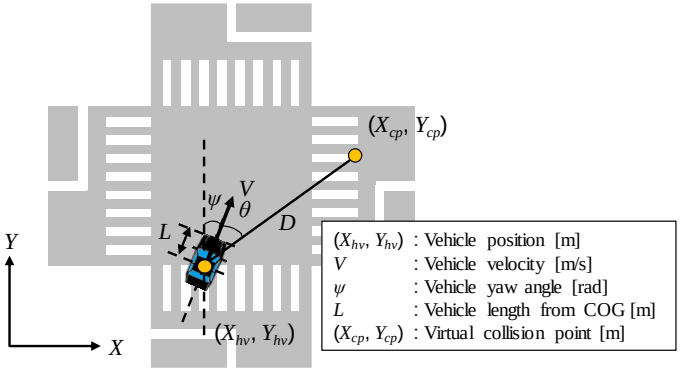


Figure 6. Model scene for TTC_{brake} calculation.

To reproduce risk scenes without putting actual pedestrians at risk, the JARI-ARV (Augmented Reality Vehicle, Uchida et al., 2015) owned by the Japan Automobile Research Institute (JARI) was used. This vehicle has video cameras and displays on its hood and the driver can see the surroundings through the displays which show the

images acquired by the video cameras. Due to this feature, this vehicle can reproduce risk scenes by superimposing computer graphic (CG) objects on real frontal images as shown in Figure 7. Using this augmented reality technology, it gives the driver the impression that the object such as vehicles and pedestrians really exist on the test field.

Figure 7. Augmented reality vehicle JARI-ARV (Top-left: Outer view of JARI-ARV, Top-right: Inner view of JARI-ARV, Bottom: Real front window image with superimposed CG objects).

Figure 8 shows the experiment course set in the test field of JARI. Non-risk scenario and risk scenario was reproduced at the target intersection with two lanes on each side and a traffic signal. The detail of the risk scenario reproduced is shown in Figure 9. Each driver drove the experiment course for a total of eight laps. Two laps for vehicle operation practice, two laps for CG scenario experience, three laps for non-risk scenario driving and last one lap for risk scenario driving.

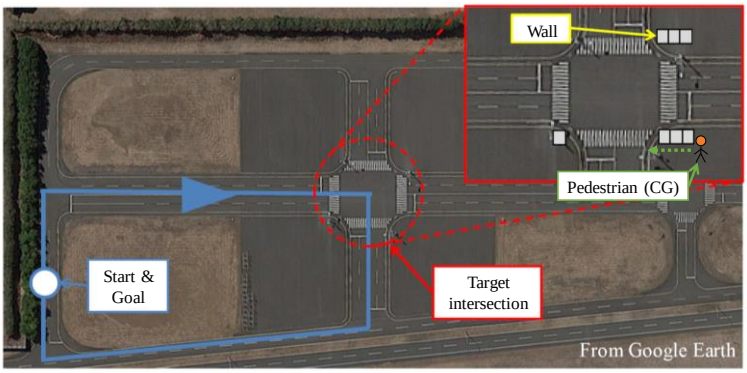


Figure 8. Experiment course.

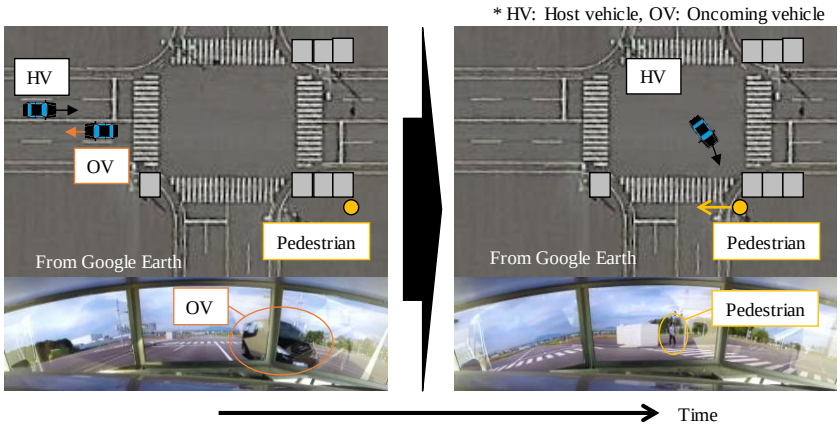


Figure 9. Risk scenario.

The subjects were 26 people (D01-D26) aged 22 to 42 years old who possessed a driving licence and drove on a daily basis. The subjects were explained about the nature of the experiment thoroughly and informed consent were obtained from each subject before the experiment.

The aim of this experiment is to reproduce risk scenes and obtain high collision risk driving behaviour data. However, it is rare to come across high collision risk incidents in daily driving. Therefore, to achieve the aim, 3 different driving conditions were set and each subject participated in the experiment with one of the conditions. The detail of each driving condition is as follows:

- Normal condition (D09-D17): Instruct subjects to drive as they do as usual.
- Hurry condition (D18-D26): Instruct subjects to drive with a hurry feeling.
- Absent-mind condition (D01-D08): Instruct subjects to drive with a secondary task (N-back task).

Result

In the risk scenario, 4 subjects did not hit the brake pedal although they noticed the crossing pedestrian and 1 subject was already pressing the brake pedal when he found the crossing pedestrian. Therefore, TTC_{brake} could not be calculated for the previously mentioned 5 subjects. Figure 10 shows the average value of TTC_{brake} for each driving condition. The variability between drivers were large and there was no significant difference between the driving conditions. From this result, although the driving condition was different between subjects, the subjects were treated as a single subject group regarding the variability as driver's characteristics.

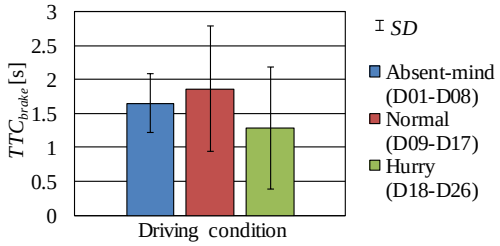


Figure 10. TTC_{brake} result of each driving condition.

Figure 11 shows the relationship between the TTC_{brake} calculated based on risk scenario driving data and the two driving behaviour indices extracted through the near-miss incident analysis calculated based on non-risk scenario driving data. The gaze duration rate (GDR) of traveling direction was calculated based on images recorded with the eye-tracking device. The analyse period was from the timing when the vehicle crossed the stop line to the timing when the driver reached the crosswalk. As the relationship shown in the figure, there were no significant correlation between the reference measure and driving behaviour indices.

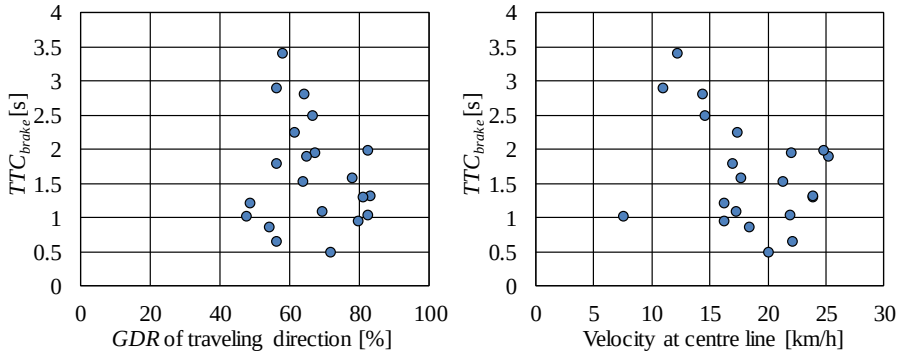


Figure 11. Driving behaviour indices vs. TTC_{brake}
(Left: GDR of traveling direction, Right: Velocity at centre line).

Although the TTC_{brake} was expected to be long when the GDR of traveling direction was low, there were cases with low GDR and short TTC_{brake}. Also, driving behaviour with high velocity at the centre line was assumed to be high collision risk behaviour but there were cases with relatively long TTC_{brake} values. These instances indicate that the two indices partially represent the collision risk with crossing pedestrians but not sufficient for collision risk evaluation. From the fact that there were cases with relatively low collision risk driving with high velocity at the centre line, the drivers could perceive the pedestrian at an early timing and hit the brake to end up with a long TTC_{brake} value despite the high vehicle speed. This suggests that there were some differences in visual search behaviour among high and low collision risk instances other than GDR of traveling direction.

Investigation of risk scenario driving behaviour

To reveal the difference of driving behaviour in high and low collision risk instances, driving behaviour data of the risk scenario was investigated. Figure 12 shows the crossing pedestrian position when the driver perceived it. The origin of the coordinate axes is the centre of gravity point of the test vehicle. The pedestrian perceived timing for each subject was set as the timing when the gaze point recorded by the eye-tracking device overlapped the crossing pedestrian in the camera image recorded. The driving behaviour data were classified into two groups (high risk driving behaviour (HR) and low risk driving behaviour (LR)) based on the median value (1.52 s). The figure shows that HR tended to find the crossing pedestrian near the vehicle front compared to LR. This result means that the drivers of HR were not able to find the pedestrian until the opponent got near to the vehicle front and this suggests that there was some difference in the visual search behaviour to find the pedestrian between HR and LR drivers.

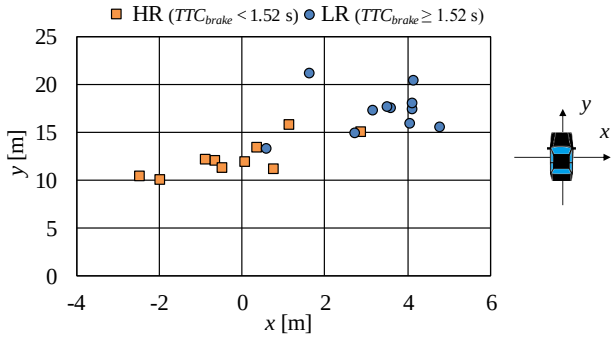


Figure 12. Position of crossing pedestrian at perceived timing.

To identify the difference, visual search behaviour of the drivers was examined. During right-turns, drivers tended to look out through the side window for some time and then started to look out through the front window as shown in Figure 13. In Figure 14, the curved line shows the vehicle path which the driver drove and the circle represents the position where the driver started to look out through the front window. Compared to the driver of LR, the visual search behaviour of the HR driver differed at the point that they looked out through the side window longer and started to look out through the front window later after getting closer to the crossing pedestrian. From the examination of driver's visual search behaviour, the timing when the driver started to look out through the front window was obtained as a feature having the possibility to evaluate pedestrian collision risk.

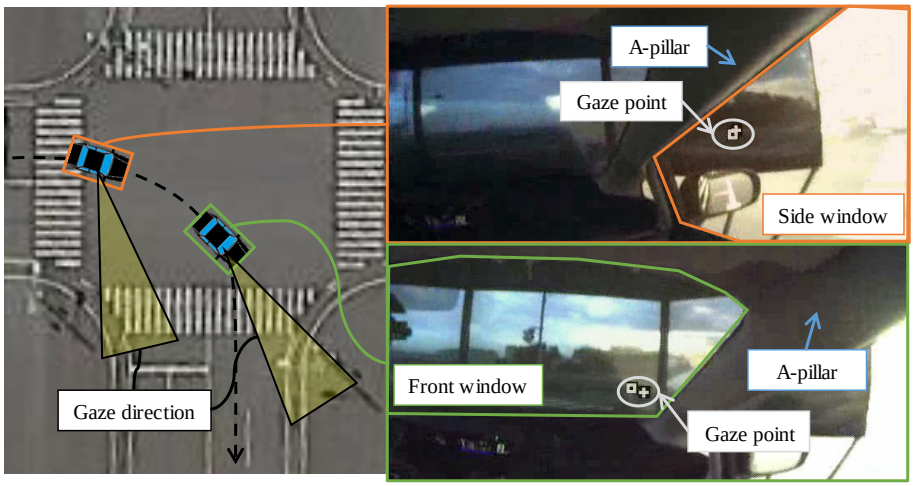


Figure 13. Visual search behaviour during right-turn at intersection.

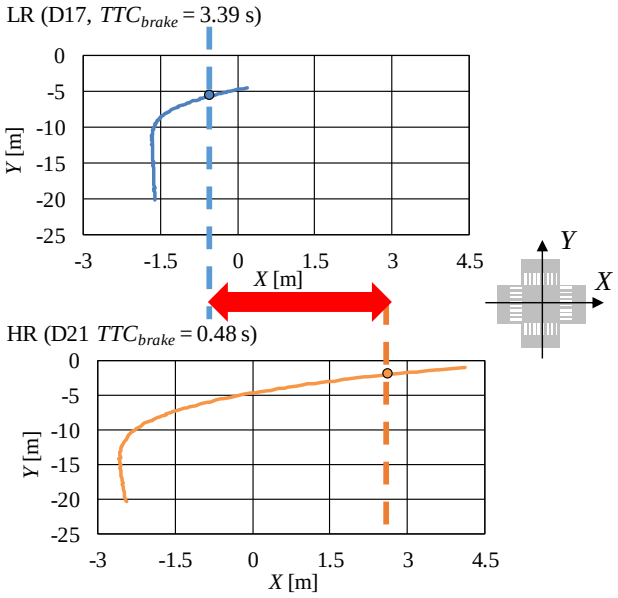


Figure 14. Comparison of vehicle position at front window confirmation timing.

To evaluate pedestrian collision risk assessment based on the previously mentioned feature of visual search behaviour, a collision risk assessment index was formulated. Because drivers look out through the side window to check the trajectory they are willing to run in the near future, it was assumed that the target trajectory geometry and the behaviour looking out from the side window have relation and consequently have relation with pedestrian collision risk. Based on the assumption that the target trajectory could be expressed by the present vehicle velocity and the change in vehicle angle per unit distance, a collision risk assessment index, estimated yaw rate

γ_{est} was defined as Equation 4 using the parameters in Figure 15. D and θ of Equation 4 was defined as Equation 5 and 6. Figure 16 shows the relationship between TTC_{brake} and estimated yaw rate γ_{est} at the centre line when the target destination point was set as the boundary of the crosswalk like it is in Figure 15. There was a significant correlation between the two indices (r = -0.65, p < .01). From this result, it was confirmed that pedestrian collision risk was evaluable by an index based on visual search behaviour of a driver.

$$\gamma_{est} = \frac{\theta}{D-L} \cdot V\cos\theta \tag{4}$$

$$D = \sqrt{(X_{dst} - X_{hv})^2 - (Y_{dst} - Y_{hv})^2} \tag{5}$$

$$\theta = tan^{-1}\left(\frac{X_{dst}-X_{hv}}{Y_{dst}-Y_{hv}}\right) - \psi \tag{6}$$

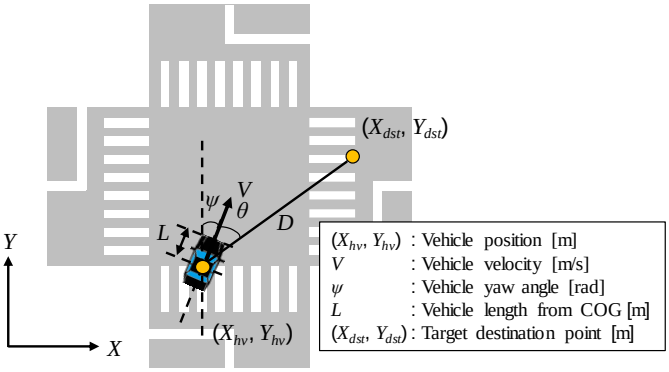


Figure 15. Model scene for γ_{est} calculation.

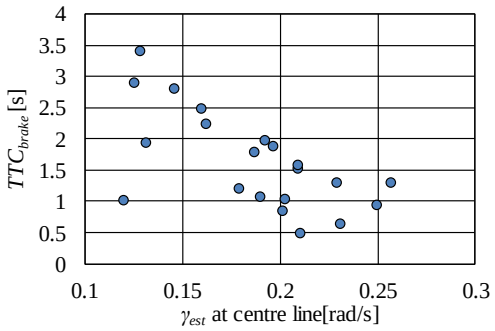


Figure 16. Relationship between TTC_{brake} and γ_{est} at centre line.

Discussion

Visual search behaviour with high pedestrian collision risk is examined by applying the visual attention modes mentioned in the introduction of this paper. Drivers tended to look out through the side window towards the traveling direction for a long time and the confirmation of the crosswalk through the front window got behind and the perception of the crossing pedestrian was late ending up with a short time-to-collision. The attention allocation towards the traveling direction can be classified as a habitual visual attention because it is an attention selection mainly for making a right-turn which is an action the driver has done hundreds of times and over-learned. The visual attention to the crosswalk is assumed to be a deliberate attention since it is driven by the risk anticipation that a pedestrian may appear. From this application of visual attention modes to high collision visual search behaviour, it can be said that shift to deliberate attention got behind because of concentration to habitual attention and pedestrian collision risk became high, as assumed in the introduction. Therefore, it is suggested that the distribution and timing of habitual and deliberate visual attention affect pedestrian collision risk.

Conclusion

To discover driving behaviour features which can evaluate pedestrian collision risk for future driver-assistance-systems, driving behaviour indices reflecting the driver's risk anticipation of a driving scene where no preceding vehicle and oncoming vehicle exists (Pattern D) was extracted based on near-miss incident analysis and the extracted indices were examined by a risk scene reproducing experiment using the JARI-ARV. The major conclusions of this research are as follows:

- The estimated yaw rate γ_{est} was verified to be a valid pedestrian collision risk assessment index for right-turns at intersections classified as Pattern D.
- Pedestrian collision risk of right-turns at intersections was evaluable by an index related to driver's visual search behaviour and it was suggested that distribution and timing of driver's habitual visual attention and deliberate visual attention, which reflects the driver's risk anticipation, affect the collision risk against pedestrians.

Acknowledgement

This research was supported by the "Next-generation advanced driver assistance system development/demonstration project" sponsored by the Ministry of Economy, Trade and Industry of Japan.

References

Engström, J., Victor, T., & Markkula, G. (2013). Attention selection and multitasking in everyday driving: A conceptual model. In M.A. Regan, J.D. Lee, and T.W. Victor (Eds.), *Driver Distraction and Inattention* (pp. 27-54). Boca Raton, Florida, US: CRC Press

Nagai, M. (2013). Present status of drive recorder database and its application potential. *Journal of Society of Automotive Engineers of Japan*, 67(2), 47-53. (in Japanese)

Shino, M., Shimazu, Y., Tagawa, T., & Kamata, M. (2015). Pedestrian collision risk indices based on driving behavior during right turns at intersections. *In FAST-zero'15 proceedings* (pp. 493-499). Gothenburg, Sweden: Chalmers University Technology

Trick, L.M., Enns, J., Mills, J., & Vavrik, J. (2004). Paying attention behind the wheel: A framework for studying the role of selective attention in driving. *Theoretical Issus in Ergonomic Science*, *5*, 385-424.

Uchida, N., Tagawa, T., & Sato, K. (2015). Development of an instrumented vehicle with Augmented Reality (AR) for driver performance evaluation. *In FAST-zero'15 proceedings* (pp. 489-492). Gothenburg, Sweden: Chalmers University Technology

Van der Hulst, M., Meijman, T., & Rothengatter, T. (1999). Anticipation and the adaptive control of safety margins in driving. *Ergonomics*, *42*, 336-345.

Modality effects of secondary tasks on hazard detection performance of younger and older pedestrians in a simulated road crossing task

Jan Siegmann, Janna Protzak, & Rebecca Wiczorek
Technische Universität Berlin
Germany

Abstract

Older pedestrians are at higher risk of being involved in car crashes than younger pedestrians. As it is known from other domains such as driving, dual-task demands represent challenges, especially for older adults. Thus, one possible reason for high accident rates of older pedestrians might be the multitasking character of the road crossing situation. With regard to safety, hazard detection represents the primary task. However, additional secondary tasks such as scanning for obstacles, navigation, and walking are mostly conducted simultaneously. According to the Multiple-Resource Model, demands of tasks regarding the main processing stage can be characterized in visual perceptual, cognitive, and motoric. The aim of the current study was to compare the effects of secondary tasks with different main process stages on the primary task of hazard detection. For this purpose, a laboratory experiment was conducted with 20 older and 20 younger participants, using a pedestrian traffic simulation. Secondary tasks differed with regard to modalities (visual search, n-back and simulated walking). Reaction time, number of errors and perceived workload served as dependent variables. Older adults performed slower, but equally accurate across all dual-task conditions compared to younger adults. Dual-task costs were found for the visual-search and the n-back task concerning number of errors, but not for response speed. No dual-task costs arouse for the walking task, which in contrast even increased hazard detection speed. The hierarchy of different dual-task costs did not differ between the two age-groups.

Introduction

Older adults (65+) accounted for 20% of injured pedestrians in Germany in 2012 (Statistisches Bundesamt, 2013). Comparing total numbers, they are not involved in accidents more often than other age groups. However, in relation to the walking distance, older pedestrians are injured more often than younger people (Rytz, 2006). Furthermore, they die more often as a consequence of the accident, as nearly half of all persons deceased after a crash were 65 years or older (Statistisches Bundesamt, 2013). That makes it important to reduce older pedestrians' crashes in traffic.

Therefore it is necessary to understand the differences of younger and older pedestrians' behaviour in road crossings and to analyse the underlying reasons.

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Several studies have found diverse behavioural deficits of older pedestrians, such as they walk slower and leave smaller safety margins, they focus more on the pathway and less on the traffic, and they do not take into account velocity but only distance information when accepting gaps in traffic (e.g. Dommes, Cavallo, Dubuisson, Tournier, & Vienne, 2014; Oxley et al., 1997; Avineri et al., 2012; Wiczorek et al., 2016). Age-related declines of cognitive and motoric functions have been identified as reasons for those differences.

However, age-related differences in specific tasks are not the only cause for older pedestrians' problems in traffic. In addition to that, one has to consider the multitasking character of the situation. Pedestrians perform several different activities during road crossing. That can be navigating, checking the pathway for bumps, and the walking itself. These activities can distract pedestrians from the main task of hazard detection and therefore be considered as secondary tasks.

While so-called dual-task costs, (performance decrements in the primary task when being engaged in a secondary task) can be observed across all age groups (Pashler, 1994), they are often higher for older than for younger adults (Verhaeghen et al., 2003). Though, a meta-analysis indicates that the extent of age-specific dual task costs depends on the characteristics of the task (Riby et al., 2004). For example, costs are higher when one of the tasks is cognitively demanding and they are more pronounced for speed than for accuracy. Furthermore it was found that older people tend to give the priority in a dual-task situation to tasks related to motion, such as walking. This effect has been referred to as 'posture first' (cf. Schaefer, 2014). The reasons may be reduced capacity to keep the balance and increased fear of falling (cf. Davidse, 2007; Dietz, 2002; Schott, 2008). Thus, the multitasking character of the road crossing situation may contribute to the occurrence of car crashes with older pedestrians.

Current studies of multitasking in road crossing usually use external tasks such as talking at the phone or texting as secondary tasks. These tasks always delay the initiation of behaviour and under some conditions reduce attentive behaviour and increase collision frequency (Neider et al., 2010; Banducci et al., 2016; Byington & Schwebel, 2013). Similar results were found when participants performed a cognitively demanding 2-back task (Gaspar et al., 2014). The only two studies that included older pedestrians found the usual age-related dual-task costs but it was mentioned that those were related to older participants' scores in motoric and cognitive screening tests (Neider et al., 2011; Nagamatsu et al., 2011). Considering the already multitasking character of the road crossing it is not surprising that older people were outperformed by younger, considering that they actually had to perform more than two tasks simultaneously. Apart from that, cell phone use in traffic is rather a problem of younger than older adults. Thus, it is necessary to investigate whether and how road crossing imbedded secondary tasks negatively affect hazard detection in younger and older adults.

With the help of the Multiple Resource Model (MRM; Wickens, 1984; 2002) it is possible to predict the amount of dual task costs dependent on the nature of the task. According to the MRM a task can be characterised along five dimensions:

- processing stage (perception vs. cognition vs. responding),
- perceptual modalities (visual vs. auditory),
- visual channels (focal vs. ambient),
- processing codes (spatial vs. verbal) and
- response execution (manual spatial vs. vocal verbal).

The more similar two tasks are on these dimensions, the higher are the predicted dual-task costs.

Research in the driving context shows the applicability of the model for the investigation of multitasking effects of older adults in traffic. For example age-specific dual-task costs were observed when participants had to respond to a secondary task manually while driving, but not when the response was given verbally (Brouwer et al., 1991; Fofanova & Vollrath, 2011). Higher costs for older adults are also found during driving when the secondary task requires visual perception, but not when it requires auditory perception (Chaparro et al., 2004; Horberry et al., 2006). Thus, it was decided to use the MRM as theoretical framework for the current investigation.

Current Study

The aim of this study is to investigate the effects of different modality specific secondary tasks on hazard detection in younger and older adults in a simulated road-crossing environment. The hazard detection task used in the current experiment consists of the perception of crossing cars in different road situations and the indication of intention to stop (i.e. not to cross the street) via joystick. Three different secondary tasks are used, which pose specifically high demands on one of the processing stages described in the MRM. The tasks require either visual-perceptual resources, cognitive resources or motoric resources.

It is expected that performance of older adults in the primary task is worse than performance of younger adults due to age-related declines in cognitive and attentional capacities (e.g. Salthouse & Somberg, 1982; Kramer & Madden, 2011; Ball, 1990). Dual-task costs should result for both age groups but should be higher for older than for younger participants, reflecting age-specific dual task cost (cf. Riby et al., 2004).

Furthermore, it is hypothesized that dual-task costs differ between the three secondary task conditions. The primary task places the highest demands in the perceptual processing stage, because potential hazards have to be detected visually. Medium demands of resources are required for the cognitive processing stage, when deciding whether the detected object represents a hazard or not. Finally only a few resources are needed for the responding stage that consists of the movement of a joystick. Thus, it is expected that dual-task costs are highest in the visual-perceptual condition followed by the cognitive condition and lowest in the motoric condition. However, as it was shown that older people give priority to motoric tasks (cf. Riby et al., 2004), the described order of dual-tasks costs is predicted only for the younger participants. For the older adults over proportionally high dual-task cost are

expected in the motoric condition and thus, the order of dual-task costs should be different for them.

Method

Participants

Forty participants of two age groups attended the study. The younger group (age ≤ 30 years) consisted of 20 students (6 male, 14 female). Their age ranged from 18 to 30 years (*M*=25.5; *SD*=3.5). The older group (age ≥ 65 years) also consisted of 20 participants (6 male and 14 female), with an age range from 67 to 82 years (*M*=71.6; *SD*=4.0). Further characteristics of younger versus older participants were respectively: possession of a driver licence (14 vs. 17), regular drivers (4 vs. 12), regular cyclists (10 vs. 9), walked regularly (20 vs. 20). Mean Montreal Cognitive Assessment (MoCA; Nasreddine et al. 2005), a screening test for mild cognitive impairment (27.65 vs. 25.75). All participants eyesight (*M*=.99 vs *M*=.63) was higher than the required minimum of 40%, (necessary to fulfil the task).

Simulation Environment and Apparatus

Animated videos are played with PsychoPy (Peirce, 2007) and projected on a wall with a width of 5,50m, a height of 1,50m, and a resolution of 3840x1080 pixels, using two Acer S1283 HNE projectors. Participants stand central at a height-adjustable standing desk 1,50m away from the projection with a visual angle of 133° to the edges of the video so that the edges of the road can only be seen through peripheral vision or head movements. Participants react to the hazard detection task using a joystick, which is attached at the top of the desk. During the cognitive and the visual-perceptual secondary task participants wore a headset. During the motoric secondary task participants used a foot switch with two pedals, which was attached at the ground under the desk. Figure 1 shows a schematic representation of the laboratory setup.

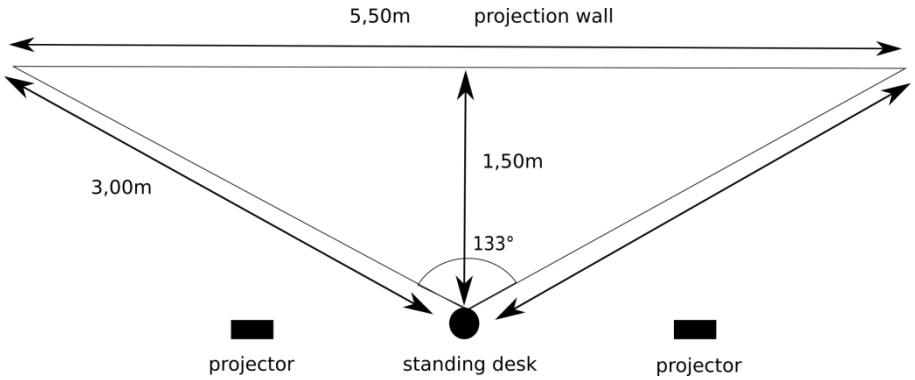


Figure 1. Schematic representation of the laboratory setup.

Hazard Detection Task

Two different types of road crossing tasks were examined in this study. One consisted of hazard detection. The other one was a ‘gap acceptance’ task and will

not be reported here. Each trial in the hazard detection task consists of a video sequence with duration of 20 seconds. In these video sequences a road is displayed and a car enters from the left or from the right edge of the scene at a pseudo random time and passes the whole screen in eight seconds. A screenshot of one of the videos can be seen in Figure 2. The road, the colour and brand of the car vary between the sequences. In total there were three different roads. Sequences of the same road were displayed in a row. Thus it was not possible for the participant to know when a trial started and thus appearance of cars was not predictable. The order of the sequences was counterbalanced between the different experimental blocks. The participants' task is to react as quickly as possible when seeing a car, by pulling the joystick towards themselves. E-bikes which are driving on the opposite pathway serve as distractors. Participants were instructed not to react to the e-bikes, which appear on average in four sequences per block in a pseudo random order.

Figure 2: Screenshot of a video used in the hazard detection task.

Secondary Tasks

Secondary tasks were created with the aim to interfere mainly with one of the three processing stages of the primary task, while not (much) interfering with the others. Therefore, the motoric and the cognitive task used the auditory perceptual modality to not interfere with the visual modality of the primary task. Cognitive demands of the motoric and the visual task were kept as low as possible to not interfere with the cognitive requirements of the primary task. The visual task and the cognitive task required vocal response execution to not interfere with the manual response of the primary task.

Visual-Perceptual Task: Participants performed a visual scanning task, which was adapted from the "Testbatterie zur Aufmerksamkeitsprüfung" (Testing Battery for Attentional Performance; Zimmermann & Fimm, 2012). A 5x6 matrix with a width of 63,5cm and a height of 52,5cm is shown centrally at the top of the projection. Every element of the matrix is a square, which is opened to one of the four sides. The target stimulus is a square opened towards the top. The task of the participants is to decide whether the matrix includes a target stimulus or not. Participants indicate their decision by saying "yes" or "no" into the microphone of the headset. Responses are recorded and a new matrix is displayed automatically after every response. Matrices alternate between white squares on black background and vice versa to make the appearance of a new matrix clear for participants. Matrices were presented in random order.

Cognitive Task: Participants performed an auditory 1-back task (Mehler et al., 2011). Every 3.33 seconds a number between zero and nine is played over a headset. After a number is read out, participants are supposed to name the number which was played one step before as quickly as possible. Responses are recorded via microphone.

Motoric Task: Participants task is to press the two pedals of a double foot switch alternating the left and the right foot in the rhythm of a regular metronome beat. Beats are played via loudspeaker every 1.33 seconds so that participants have to press the switch 45 times per minute. This frequency corresponds to a cadence of 90 steps per minute (pressing the foot switch requires two steps: moving the foot forward to the pedal and backward again). The cadence of most healthy elderly adults lies between 80 and 130 steps per minute (Whittle, 2014).

Design and Dependent Measures

Design: The study consisted of a 2x4 mixed design. The between factor was age-group and the within factor was secondary task with four conditions (baseline/single task, cognitive dual-task, visual-perceptual dual-task, and motoric dual-task. The dependent measures were:

- Number of errors: sum of false negatives and false positives
- Reaction time: between time of appearance of a car and pull of the joystick. (Reaction time was assessed via frame numbers of the video)
- Subjectively perceived workload: assesses with the NASA- Task Load Index (NASA-TLX ; Hart & Staveland, 1988)

Procedure

On arrival, participants were briefly instructed about the course of the experiment and filled in a declaration of consent. Afterwards they performed an eyesight test with Landolt rings. Following up participants were acquainted with the laboratory equipment. Participants trained each of the secondary tasks and performed it for five minutes in order to get familiar with the task. The order of the tasks was counterbalanced. After a 5-minute break participants trained the two primary tasks (hazard detection and gap acceptance). The experiment started with the first baseline measure block which consisted of five minutes single task hazard detection and five minutes single task gap acceptance. The order of the tasks was counterbalanced between participants. Subsequently participants completed three dual-task blocks with a duration of 2x5 minutes each. The order of the secondary tasks and the road crossing tasks corresponded to the order of the training. Participants had been instructed that both tasks were important. The secondary tasks started 20 seconds before the 5-minute road crossing tasks. The first and the last dual-task block were followed by a 5-minute break. Afterwards participants performed a second baseline block. The average of the measures of both baseline blocks was used for the analysis in order to prevent biases due to learning effects or fatigue. After each of the baseline and experimental blocks participants filled in the NASA-TLX. In the end of the experiment participants filled in the MoCA and a demographic questionnaire. Finally they received a financial compensation.

Results

To test for effects of learning and fatigue the data of the two baseline measures was compared using a 2x2 ANOVA for repeated measures. Afterwards the average of the two baseline measures was used for further calculations. The number of errors, reaction time and subjective workload were compared between dual-task conditions using 2x4 ANOVAs for repeated measures. A priori defined contrasts were used to analyse the order of dual-task costs between the three secondary tasks.

Comparison of Baseline blocks: Analysis revealed a main effect of order, $F(1,38)=5.109$, $p<.05$, and a main effect of age, $F(1,38)=5.630$, $p<.05$, which were further qualified by a significant age x order interaction effect, $F(1,38)=7.983$, $p<.001$. Older participants made more errors than younger people in the first baseline block but reduced there number of errors to the lower level of the younger group in the second baseline. This indicates a learning effect. No significant differences were found for reaction time and subjective workload between the two age groups and the two measures.

Error rate: Analysis revealed a main effect of secondary task, $F(3,114)=9.737$, $p<.001$, but no other significant effects. A priori defined contrast showed significant differences between the visual-perceptual condition and the baseline, $F(1,38)=12.507$, $p<.001$, as well as the motoric condition, $F(1,38)=13.760$, $p<.001$. The same pattern revealed for the cognitive condition when compared to the baseline, $F(1,38)=16.130$, $p<.001$ and the motoric condition, $F(1,38)=16.369$, $p<.001$. As can be seen in Figure 3, more errors were made in the visual-perceptual condition and in the cognitive condition, compared to the baseline and the motoric condition.

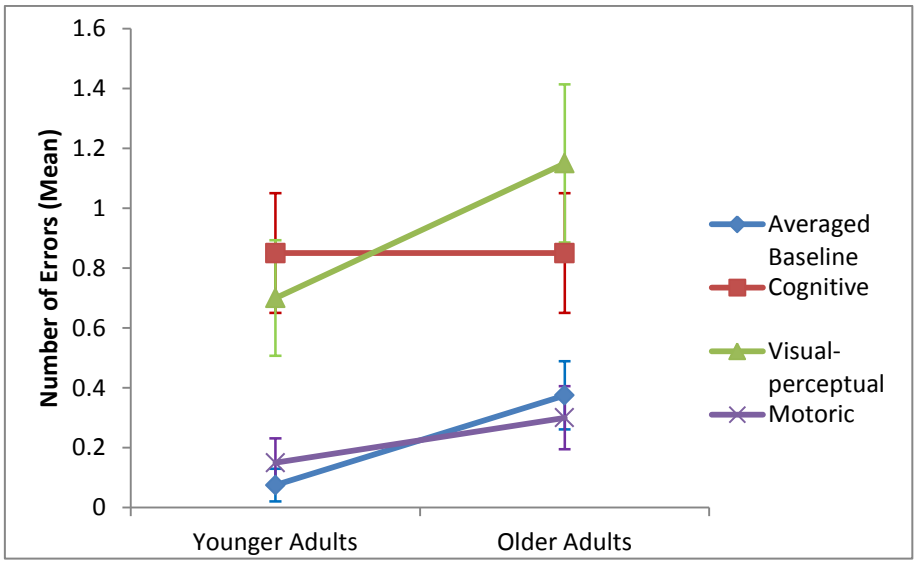


Figure 1: Means of number of errors of the averaged baseline and the three dual-task conditions for younger and older adults. Error bars reflect Standard Error.

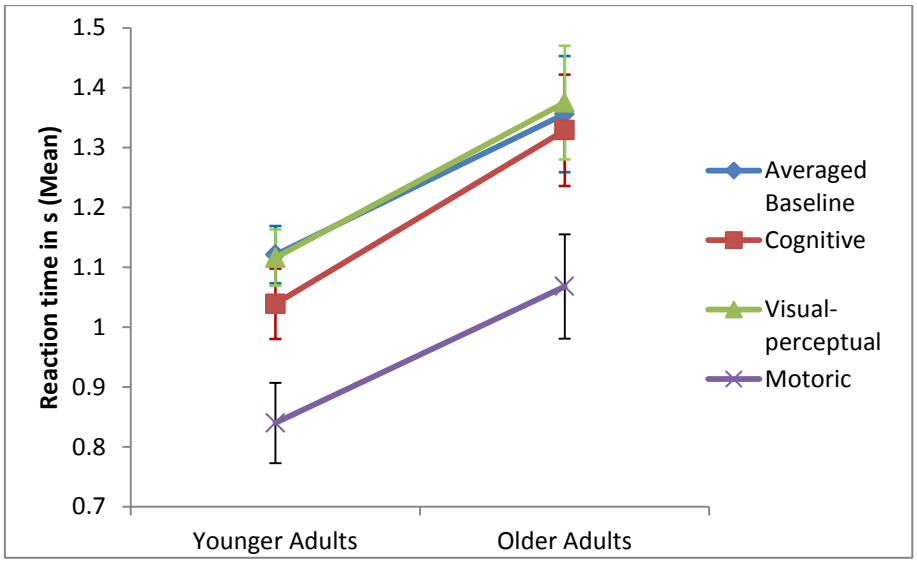


Figure 4: Means of reaction time of the averaged baseline and the three dual-task conditions for younger and older adults. Error bars reflect Standard Error.

Reaction time: The concurrent presentation of matrices and videos in the visual-perceptual dual-task condition led to a reduction of M=10,8 frames per matrix, which had to be corrected manually afterwards. Comparison of the averaged baseline measure and the three dual-task conditions showed a main effect of age, $F(1,38)=6.431$, $p<.05$ with older adults being slower than younger adults. Furthermore a main effect of secondary task was observed, $F(1,38)=31.093$, $p<.001$). Planned contrasts showed that the reaction time in the motoric condition was shorter than in the baseline $F(1,38)=69.299$, $p<.001$. The reaction time in the motoric condition was also shorter than in the cognitive, $F(1,38)=36.878$, $p<.001$, and the visual-perceptual condition, $F(1,38)=55.646$, $p<.001$. Reaction times between the baseline and the cognitive and the visual-perceptual condition did not differ. Figure 4 shows results of reaction time.

Subjective workload: Comparison of the data from the NASA-TLX revealed a main effect of secondary task, $F(3,114)=60.499$, $p<.001$. Planned contrasts showed the difference of perceived workload between all four conditions with the least workload in the baseline condition, followed by the motoric condition, $F(1,38)=65.548$, $p<.001$, the visual-perceptual condition, $F(1,38)=13.508$, $p<.001$, and the highest workload in the cognitive condition, $F(1,38)=5.275$, $p=.027$. F-values refer to contrasts with the condition with the next lower workload. No main effect of age and no interaction effect were found. Results can be seen in Figure 5.

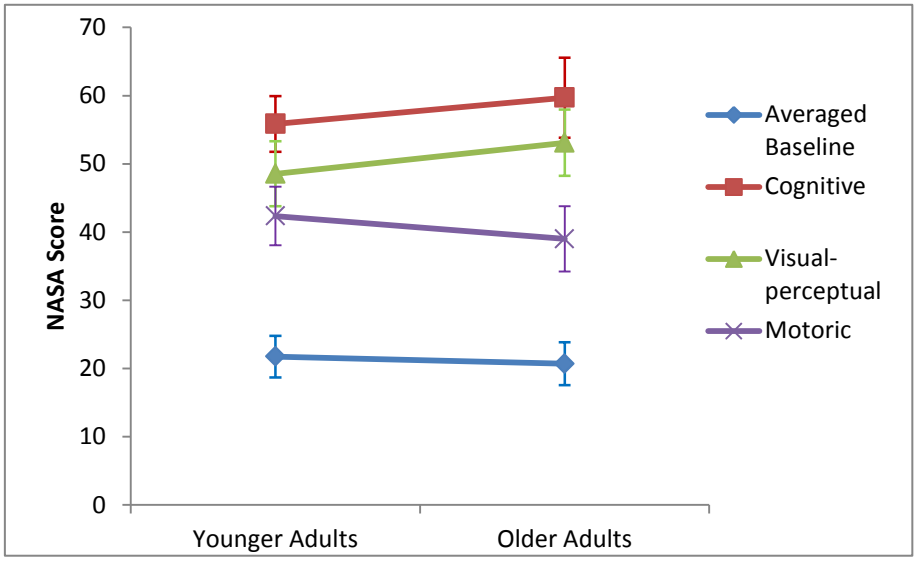


Figure 5: Means of reaction time of the averaged baseline and the three dual-task conditions for younger and older adults. Error bars reflect Standard Error.

Discussion

The current study was conducted to investigate the effects of road crossing imbedded secondary tasks on hazard detection performance in simulated pedestrian crossing situations. The aim of this study was to identify differences and similarities of younger and older pedestrians when performing secondary tasks, which differed with regard to their main processing stage. Number of errors, reaction time and subjective workload were assessed. Dual-task related performance decreases in the primary task of hazard detection were expected for all the participants. The older adults however, should additionally produce age-specific (i.e. higher) dual-task costs than the younger group. It was also hypothesized that the three secondary tasks produce different amounts of dual-task costs. The highest costs were expected for the visual-perceptual secondary task, followed by the cognitive task, with the fewest for the motoric task. This order is based on the different demands of the primary task on the three processing stages. Hazard detection should require the most resources for the visual-perceptual processing stage, followed by the cognitive stage and only a minimum of motoric resources as the motoric component consisted of moving a joystick. However, it was further hypothesised that this order would be different for older adults. As they were found to not fulfil the task of walking as automatically as younger people and to need additional attention resources, it is possible that the motoric task also interferes with one of the other processing stages.

The comparison of two baseline measures conducted at the beginning and at the end of the experiment indicated a learning effect of older adults. For the following analysis, the two baseline blocks were averaged.

No difference between younger and older adults could be observed with respect to the number of errors. However, older adults reacted slower than younger adults. This is in line with previous research, which found higher age-specific performance decrements when considering response speed than when considering response accuracy (Verhaeghen et al., 2003).

Dual-task costs were observed regarding number of errors, but not for reaction time. Older and younger adults made more errors in the primary task of hazard detection when they performed a cognitive or visual-perceptual task concurrently. It seems that participants accepted making more errors in order to maintain their reaction speed. In real-life road crossing this behaviour can have fatal consequences, when pedestrians overlook potential hazards. However, it has to be noted that the number of false positives (reactions in the absence of cars, for example to the e-bikes used as distractors) was higher than the number of misses. Nevertheless also false positives can lead to dangerous traffic situations, for example when pedestrian behaviour diverts from the expectations of other road users

In contrast to the other two secondary tasks, the motoric task not only did not cause any dual-task costs, but even improved performance in terms of reaction speed. There are some possible explanations for this unexpected result. First, the motoric task might have been too easy, compared to normal walking. Participants could hold on the standing desk and had always one food on the ground. Thus, there was no risk of losing balance, which can explain why no dual-task costs were found. Furthermore, neuropsychological research shows that motion can influence visual perception, either directly or via attentional mechanisms (Ishimura & Shimojo, 1994). That could explain the observed benefits for the hazard detection when conducting the motoric task in parallel. In an experiment with mice, Ayaz et al. (2013) found neurons in the visual cortex to have a higher firing frequency when subjects were running on a treadmill. If this effect holds true also for humans, it could explain the faster reaction times in the motoric condition.

The unexpected results in the motoric condition could also explain why the order of dual-task costs did not differ between younger and older adults. It was expected that the motoric task would only produce small dual-task costs in younger adults, because the primary task did only induce minor motoric demands. However, older adults were expected to suffer from higher dual-task costs, because for them walking is less automatic and requires more attentional resources (Lindenberger et al., 2000). Thus, walking could possibly interfere also with the visual-perceptual and the cognitive task. In order to examine this hypothesis a more realistic motoric task should be used in further research.

However, it has to be noted that also the subjectively perceived workload did not differ between younger and older adults but showed significant differences among all four conditions. The baseline single task condition was considered the least demanding followed by the motoric dual-task, the visual-perceptual and the cognitive dual-task. Differences in the NASA-TLX score for the visual-perceptual and the cognitive dual-task condition were not accompanied by differences in performance. Performance in the single task condition was similar to the visual-

perceptual and the cognitive dual-task conditions with regard to number of errors but resembled most the motoric dual-task in terms of reaction time.

The three secondary tasks were created in a way they would interfere with only one of the three processing stages according to the MRM. As expected the least impairment was observed in the motoric process, which only consisted of pulling the joystick in the primary task and therefore did not require a lot of resources. No significant differences were found between the visual-perceptual and the cognitive task with regard to number of errors and reaction time. However, it is possible that differences would emerge on a behavioural level rather than on the performance level. As it was shown that scanning behaviour in road crossing differs between older and younger adults (Tapiro et al., 2016), investigation of eye movements could possibly also show differences between secondary tasks.

Findings of this study can be used to make road crossing safer for younger as well as for older adults. They underline the need for awareness campaigns, which point out the risks of multitasking during road crossing. Results also show that crossing points should be designed in a way that they induce as few additional workload as possible. This can be achieved by using smooth sidewalks and minimising billboards, especially on places which are used frequently for road crossing. Findings of this study will also help to clarify demands of a pedestrian assistance system for older adults, which is currently developed by the junior research group FANS at the TU Berlin (cf. Breitinger et al., 2015). According to current findings, the system should support users to investigate all resources in hazard detection. That means, execution of other road crossing imbedded tasks such as scanning the ground for obstacles and navigation should be done before or after but not in parallel to the primary task. Whether the same holds true for concurrent walking will be further analyses using a more realistic walking task.

References

Avineri, E., Shinar, D., & Susilo, Y.O. (2012). Pedestrians' behaviour in cross walks: The effects of fear of falling and age. *Accident Analysis & Prevention*, *44*, 30–34.

Ayaz, A., Saleem, A.B., Schölvinck, M.L. & Carandini, M. (2013). Locomotion controls spatial integration in mouse visual cortex. *Current Biology*, *23*, 890–894.

Ball, K.K., Roenker, D.L., & Bruni, J.R. (1990). Developmental changes in attention and visual search throughout adulthood. *Advances in psychology, 69*, 489-508.

Banducci, S.E., Ward, N., Gaspar, J.G., Schab, K.R., Crowell, J.A., Kaczmarski, H., & Kramer, A.F. (2016). The effects of cell phone and text message conversations on simulated street crossing. *Human Factors*, *58*, 150–162.

Byington, K. W., & Schwebel, D. C. (2013). Effects of mobile Internet use on college student pedestrian injury risk. *Accident Analysis & Prevention*, *51*, 78–83.

Breitinger, F., Protzak, J. & Wiczorek, R. (2014). Conceptualizing everyday mobility of older people as basis for the development of a pedestrian assistance system. *Studies in Health Technology and Informatics, 217*, 935–940.

Brouwer, W.H., Waterink, W., Van Wolffelaar, P.C., & Rothengatter, T. (1991). Divided attention in experienced young and older drivers: lane tracking and visual analysis in a dynamic driving simulator. *Human Factors*, *33*, 573–582.

Chaparro, A., Wood, J.M., & Carberry, T. (2004). Effects of age and auditory and visual dual-Tasks on closed road driving performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*, 2319–2322.

Davidse, R. (2007). *Assisting the older driver*. [dissertation]. University of Groningen, Groningen, Netherlands. http://irs.ub.rug.nl/ppn/30535423X.

Dietz, V. (2002). Proprioception and locomotor disorders. *Nature Review Neuroscience*, 3(10), 781–790.

Dommes, A., Cavallo, V., Dubuisson, J.-B., Tournier, I., & Vienne, F. (2014). Crossing a two-way street: comparison of young and old pedestrians. *Journal of Safety Research*, *50*, 27–34.

Fofanova, J., & Vollrath, M. (2011). Distraction while driving: The case of older drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*, 638–648.

Gaspar, J.G., Neider, M.B., Crowell, J.A., Lutz, A., Kaczmarski, H., & Kramer, A.F. (2014). Are gamers better crossers? An Examination of action video game experience and dual task effects in a simulated street crossing task. *Human Factors*, *56*, 443–452.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P.A. Hanock & N. Meshkati (Eds.), *Advances in Psychology* (pp. 139–183). Amsterdam: North-Holland.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention*, *38*, 185–191.

Ishimura, G. & Shimojo, S. (1994). Voluntary action captures visual-motion. *Investigative Ophthalmology & Visual Science*, *35*, 1275–1275.

Kramer, A.F., & Madden, D.J. (2011). Attention. In F.I.M. Craik and T.A. Salthouse (Eds.), *The Handbook of Aging and Cognition: Third Edition* (pp. 189–250). New York: Psychology Press.

Lindenberger, U., Marsiske, M. & Baltes, P.B. (2000). Memorizing while walking: Increase in dual-task costs from young adulthood to old age. *Psychology and Aging*, *15*, 417–436.

Mehler, B., Reimer, B. & Dusek, J.A. (2011). *MIT AgeLab delayed digit recall task (n-back)* (MIT AgeLab White Paper No. 2011-3B). Cambridge: Massachusetts Institute of Technology.

Nagamatsu, L.S., Voss, M., Neider, M.B., Gaspar, J.G., Handy, T.C., Kramer, A.F., & Liu-Ambrose, T.Y. (2011). Increased cognitive load leads to impaired mobility decisions in seniors at risk for falls. *Psychology and Aging*, *26*, 253-259.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*, 695–699.

Neider, M.B., Gaspar, J.G., McCarley, J.S., Crowell, J.A., Kaczmarski, H., & Kramer, A.F. (2011). Walking and talking: Dual-task effects on street crossing behavior in older adults. *Psychology and Aging*, *26*, 260–268.

Neider, M.B., McCarley, J.S., Crowell, J.A., Kaczmarski, H., & Kramer, A.F. (2010). Pedestrians, vehicles, and cell phones. *Accident Analysis & Prevention*, *42*, 589–594.

Oxley, J., Fildes, B., Ihsen, E., Charlton, J., & Day, R. (1997). Differences in traffic judgements between young and old adult pedestrians. *Accident Analysis & Prevention*, *29*, 839–847.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244.

Peirce, J.W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13.

Riby, L.M., Perfect, T.J., & Stollery, B.T. (2004). The effects of age and task domain on dual task performance: A meta-analysis. *European Journal of Cognitive Psychology*, *16*, 863–891.

Rytz, M. (2006). *Senioren und Verkehrssicherheit* [Seniors and traffic safety]. Bern: VCS Verkehrs-Club der Schweiz.

Salthouse, T.A., & Somberg, B.L. (1982). Isolating the age deficit in speeded performance. *Journal of Gerontology*, *37*, 59–63.

Schaefer, S. (2014). The ecological approach to cognitive–motor dual-tasking: findings on the effects of expertise and age. *Frontiers in psychology*, 5, 1167.

Schott, N. (2008). Deutsche Adaptation der "Activities-Specific Balance Confidence (ABC) Scale" zur Erfassung der sturzassoziierten Selbstwirksamkeit. [German adaptation of the "Activities-Specific Balance Confidence (ABC) Scale"]. *Zeitschrift für Gerontologie und Geriatrie*, 41, 475-485.

Statistisches Bundesamt (2013). *Unfallentwicklung auf Deutschen Strassen 2012* [Accident development on German roads in 2012]. Begleitmaterial zur Pressekonferenz am 10.Juli 2012 in Berlin. https://www.destatis.de/DE/Publikationen/

Tapiro, H., Borowsky, A., Oron-Gilad, T., & Parmet, Y. (2016). Where do older pedestrians glance before deciding to cross a simulated two-lane road? A pedestrian simulator paradigm. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 11-15). SAGE Publications.

Verhaeghen, P., Steitz, D.W., Sliwinski, M. J., & Cerella, J. (2003). Aging and dual-task performance: a meta-analysis. *Psychology and Aging*, *18*, 443.

Whittle, M.W. (2014). *Gait analysis: An introduction*. Oxford: Butterworth-Heinemann.

Wickens, C.D. (1984). Processing resources and attention. In R. Parasuraman & D.R. Davies (Eds.), *Varieties of Attention.* (pp. 63–102). New York: Academic Press.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*, 159–177.

Wiczorek, R., Siegmann, J., & Breitinger, F. (2016). Investigating the impact of attentional declines on road-crossing strategies of older pedestrians. In D. de Waard, K.A. Brookhuis, A. Toffetti, A. Stuiver, C. Weikert, D. Coelho, D. Manzey, A.B. Ünal, S. Röttger, and N. Merat (Eds.), *Proceedings of the*

Human Factors and Ergonomics Society Europe Chapter 2015 Annual Conference (pp. 155-169). Available from http://hfes-europe.org

Zimmermann, P. & Fimm, B. (2012). *Testbatterie zur Aufmerksamkeitsprüfung: Version Mobilität (Manual) [Testing battery for attentional performance: Version mobility (manual)]*. Herzogenrath: Psytest.

Impact of interface sonification with touchless gesture command in a car

Ludovic Jaschinski[1], Sébastien Denjean[2], Jean-François Petiot[1], Frank Mars[1], & Vincent Roussarie[2]
[1]IRCCyN, UMR CNRS 6597, Ecole Centrale de Nantes, [2]PSA Peugeot Citroën France

Abstract

This experiment aims to study the impact of the sonification of a hand gesture controlled system on the driver behavior. The principle is to give an auditory feedback to the driver, in addition to a visual screen, in order to assist in-car devices interface manipulation. A touchless interface has been tested with a panel of 24 subjects on a driving simulator. Different tasks (pick up the phone, select an item in a list) involving the screen and the interface had to be performed by the user while driving. To study the contribution of the sound feedback on the drivers' behavior, two audio conditions were tested: with and without auditory feedback. The tasks were performed in lowly and highly demanding traffic conditions. Driving and gaze behavior as well as eye-tracking information were analyzed. Moreover, a questionnaire was used to obtain subjective measurements, such as ease of use and feeling of safety. The results show that the sonification helped drivers to feel safer and more focused on the road. This result was confirmed by gaze analysis, which shows that drivers look significantly less to the visual interface when a sound is present, leading to a safer use of the interface.

Introduction

The manipulation of in-vehicle information systems is a challenge in today's vehicle design. These systems are more and more interactive while their complexity is expanding, leading to *infotainement* systems that include many functions. When the user is engaged in a primary task (driving), the basic solution that involves a physical interaction with the device while looking at a visual display is an interaction way that can certainly be improved. In particular, auditory feedbacks may decrease the need for visual attention and free-hand gesture interaction eliminates the need for reaching the device. The sonification of information is an emerging discipline that exploits the capacity of sounds to convey information (Hermann et al., 2011) and many applications are proposed in particular in the car industry (Denjean et al., 2013). In this context, touchless interfaces with auditory feedback may be an interesting alternative to assist and even replace visual interfaces. (Kajastila & Lokki, 2013) showed that auditory interfaces can outpoint visual interfaces, in particular with free-hand (touchless) interaction. A study of multi-modal controls (visual, audio or visual+audio) of in-vehicle information

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

systems is presented in (Jakus et al., 2015) with a driving simulator. The results showed that the visual and visual+audio interaction modes are the most fast and efficient, but no significant contribution of the audio to the visual-only mode on the driving performance and safety is noticed: the main contribution of audio is that it increased drivers' preferences. (Sodnik et al., 2008) presented a comparative study on the effectiveness and efficiency of audio-only or visual-only interfaces for the manipulation of a mobile phone while driving on a simulator. They demonstrated that audio-only interfaces were effective to control a mobile, particularly when spatialized auditory cues are used in the audio interface. A study with a real vehicle on the acceptability of a sonification for GPS navigation (Tardieu et al., 2015) showed that with sounds, the driving was estimated as safer by the drivers, even if it did not improve significantly the efficiency of the navigation.

In this context, our work aims to assess the contribution of sounds to a visual touchless interface. The objective is to study the impact of the interface sonification on the driver behavior, during the manipulation of an infotainement system in a car. An experiment was conducted on a driving simulator, with two main experimental conditions: manipulation of the interface with visual-only interaction, or with visual+audio interaction. Two categories of variables were observed: driving parameters, recorded by the simulator, but also the driver gaze behavior, recorded by means of an eyetracking system. After a presentation of the material and methods of the experiment, the results are presented and discussed.

Material and methods

Interface

A graphical interface with gestures command without contact (using a *Leap motion*[1] device) was implemented. Six functional categories were created to constitute the main menu of the infotainement system (phone, air conditioning, contacts, music, news and GPS). The main menu of the interface is given in figure 1.

Figure 2. Main menu of the interface of the in-vehicle information system (French version)

[1] https://www.leapmotion.com

Different contactless gestures, necessary to control the interface, were programmed:

- *Selection* of a category in the main menu: the subject has to point with one finger (index finger) which category he/she wants to select (selection gesture)
- *Validation*: this is made by closing the fist (validation gesture)
- *Browsing* into a list of items (draggable carousel menu): this is done by sweeping the items on the right or left side (horizontal movement of the hand on the right or left) (sweeping gesture)
- *Back return* in the menu: this is done by rotating the hand the palm facing up (return gesture)

The general framework of the interface is given in figure 2. It allows a real time control of the screen and of the played audio samples, according to the gestures provided by the subject:

- The leap motion detects subject' gestures and sends the information to a general program (coded in python).
- The program updates the interface screen in real time and sends informations to the sound synthesizer (*Pure Data*[2] code).
- The syntheziser sends audio samples to the loudspeakers.

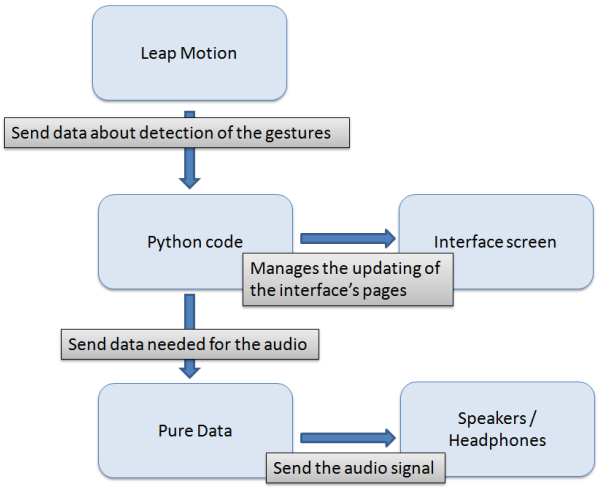


Figure 3. Framework of the interface for the realtime control of the screen and the audio samples

[2] https://puredata.info : Pure Data (aka Pd) is an open source visual programming language that processes and generates sounds based on a graphical interface.

Sonification

Different sounds were associated to the gestures of the user (sonification) in order to provide an audio feedback and help the handling of the interface. The sounds were designed from samples and sounds synthesized on the digital audio workstation *Reaper*[3], with MIDI notes. The following 11 sounds have been designed:

- 6 sounds corresponding to the 6 categories on the main menu, to specify the area pointed by the hand (selection gesture),
- a validation sound (validation gesture)
- a sweep / scrolling sound (sweeping gesture for browsing)
- a sweep / validation sound (sweeping gesture for validation)
- a back return sound (return gesture)
- a refusal sound (pop-up) (sweeping gesture for refusal)

All the sounds were very short (between 50 and 300ms) and they were designed to be easily recognized by the user.

Experiment

Material
The experience took place on the driving simulator of the IRCCyN laboratory, shown in figure 3.

Figure 4. Picture of the fixed-base simulator used for the driving tests

It is a fixed-base simulator, which consist of a compact size passenger car with actual instrument panel, clutch, brake and accelerator pedals, handbrake, ignition key, and an adjustable seat with seat belt. The visual environment was displayed on three 32-inch LCD monitors, each with a resolution of 1280□*720. One monitor was positioned in front of the driver, with two laterals inclined at 45 deg from the

[3] http://www.reaper.fm. Audio processing software

front one, viewed from a distance of about 1 m and covering 115 deg of visual angle. An additional screen was added to the simulator for the interface (figure 4).

Figure 5. Picture of the screen to display the interface and of the leapmotion to capture the gestures of the subject

The Leap Motion was placed in front of this additional screen and connected to the computer of the interface. Speakers were placed behind the simulator screens to play the sound samples. A Smarteye Pro 5 eye-tracking system, composed of 4 cameras placed under the three screens in front of the subject, was calibrated to measure the location and duration of gaze fixations (glances) during the tasks.

Subject and experimental factors
24 subjects (16 men and 8 women – average age 24), students or researchers at Ecole Centrale de Nantes, participated to the tests. The subjects had to drive on a countryside road, on the same route for each scenario. The experimental factor was the sonification condition, with two levels: with sound (s) and without sounds (w-s). For each sonification condition, every course was carried out in two different conditions: with no visible danger on the road (free-flowing traffic) and with a disturbing traffic (a vehicle in front of the driver that drives slowly and brakes regularly). The later case should yield an increased mental workload (not measured). Four scenarios are thus envisaged:

- Scenario 1: no sound, free-flowing traffic
- Scenario 2: no sound, disturbing traffic
- Scenario 3: with sound, free-flowing traffic
- Scenario 4: with sound, disturbing traffic

Interface manipulation tasks
For each course of the scenario, the following four tasks had to be completed by the subject on the interface, in a straight line and in a curve:

- task 1: Acknowledge receipt of a message (pop-up), in a straight line

- task 2: Select a particular destination on the GPS, in a curve
- task 3: Select a particular destination on the GPS, in a straight line
- task 4: Acknowledge receipt of a phone call (pop-up), in a curve

Tasks 1 and 4 are simple: they require only one gesture from the subject (sweeping gesture). Tasks 3 and 4 are more complex: they require several different gestures (selection, browsing, validation, back return). The trigger of the tasks and the corresponding instructions were shown to the subject on the central screen of the simulator. The tasks had to be made always on the same location for every trial. The map of the course and the positioning of the tasks are presented in figure 5.

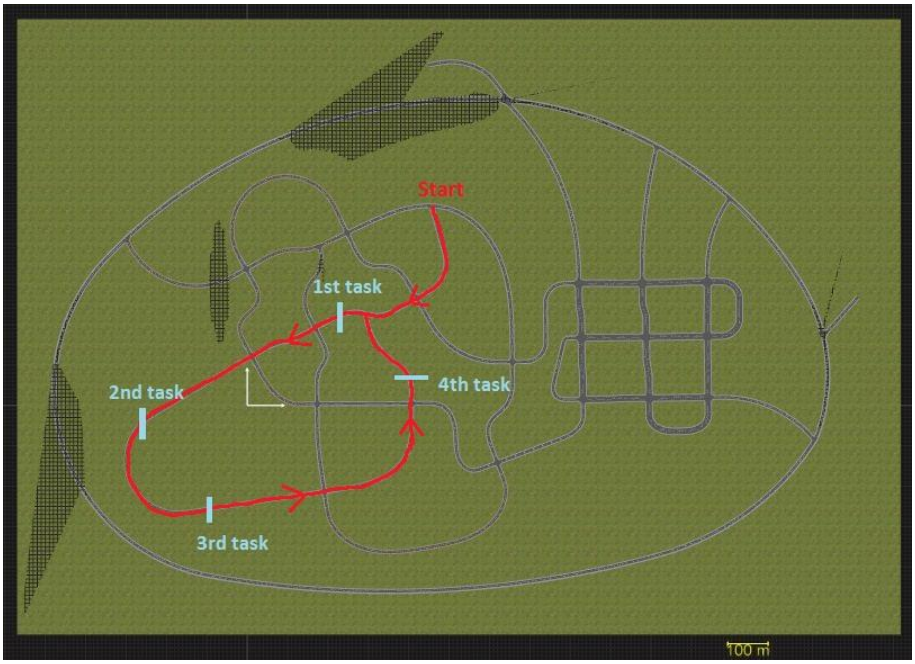


Figure 6. Map of the course and location of the four tasks

Test procedure
First, a training period was proposed to the participants so that they get acquainted with the *Leap Motion* and with the production of the gestures, with the manipulation of the interface, and with the simulator. No subject was familiar with the use of the *Leap motion*. For the interface, the subjects were trained without sound first, and then with sounds. Then, every subject made each of the four scenarios in a counterbalanced order.

Dependant variables (DV)
The following dependent variables were considered, for each subject, each task and each scenario, between the trigger of the task and its completion by the subject:
Driving variables (9 variables):

- Completion time of the task,

- Average and standard deviation (SD) of the lateral deviation of the vehicle,
- Number of steering wheel reversal,
- Average and standard deviation (SD) of the speed of the vehicle,
- Average and standard deviation of longitudinal acceleration,
- Inter-vehicular time.

Eye tracking variables (5 variables):

- Glance duration on the interface screen and number of glances,
- Glance rate (corresponding to the glance duration divided by the time for performing the task) on the interface screen,
- Maximum time of glance duration on the interface screen (without looking back on the simulator screens),
- Number of eye movements from the simulator screens to the interface screen.

Questionnaire

A semi-directive questionnaire was proposed to the participants after the completion of each scenario. Different questions were formulated to assess:

- the global impression: ease of use, feeling of safety, feeling of not looking at the road,
- the impression on the sound design: is it disturbing, is it appreciated?
- two open questions, one on the global interface and one on the sound design

Data analysis

Descriptive statistics on the DV showed that few spurious data were present in the recordings. To reduce the effect of these outliers on the DV, a winsorisation was applied on each DV (extreme values above the 95^{th} percentile and below the 5^{th} percentile were setted to the 95^{th} and 5^{th} percentile respectively) (Wilcox, 2014). Each dependent variable (DV) was next analysed with ANOVA (Næs et al., 2010). Two ANOVA were carried out:

- one-way ANOVA with the factor “traffic condition” (two levels: free-flowing traffic; disturbing traffic)
- two-ways ANOVA with interaction with the factors “subject” (24 levels) and “sonification condition” (two levels: with sounds (s) and without sounds (w-s))

The significance of the effects was analysed with the Fisher test (type III sum of square) and the associated p-value (significance level: $p<.05$)

Results and discussion

Effect of the factor “traffic conditions” (one way ANOVA)

The results show that, for all the DVs considered, there is no significant effect of the “traffic conditions” (one way ANOVA, $p>.05$). The subjects were very focused on

the road even when the traffic condition was not difficult because they wanted to respect as well as possible the driving instructions. For this reason, this factor is ignored in the following analyses.

Effects of the factors "subject" and "sonification conditions" (two-ways ANOVA)

The results show that the factor "subject" was significant ($p<.05$) for most of the DVs (detailed results are not reported here for concision). This is a sign of singificant inter-individual differences in the driving performances and in the management of the interface. This result was expected and does not need particular comments. The interactions subject*sonification was almost never significant.

Table 1 presents the results of the two-way ANOVAs for the factor "sonification condition", for the two categories of DVs (Driving and Eye tracking). When significant, the sign of the effect (difference between the level with sound (s) and without sound (w-s)) is mentioned with the relation w-s>s or w-s<s).

*Table 1. Results of ANOVAs: significance of the "sonification condition" for the different dependent variables (without sound: w-s – with sound: s). F-test: * p<.05 - **p<.01 – n.s. : not significant (p>.05)*

	Dependent variable DV	*Task 1*	*Task 2*	*Task 3*	*Task 4*
Driving variables	Completion time	n.s.	n.s.	n.s.	n.s.
	Average lateral deviation	n.s.	n.s.	n.s.	n.s.
	SD of lateral deviation	** w-s > s	n.s.	n.s.	n.s.
	Number of steering wheel reversals	n.s.	n.s.	n.s.	n.s.
	Average speed	n.s.	n.s.	n.s.	n.s.
	SD of speed	n.s.	n.s.	n.s.	n.s.
	Average longitudinal acceleration	** w-s < s	n.s.	n.s.	n.s.
	SD of longitudinal acceleration	n.s.	n.s.	** w-s > s	** w-s > s
	Inter-vehicular time	n.s.	n.s.	n.s.	** w-s < s
Eye tracking variables	Glance duration on the interface screen	n.s.	** w-s > s	** w-s > s	n.s.
	Number of glances	n.s.	n.s.	n.s.	n.s.
	Glance rate	n.s.	** w-s > s	** w-s > s	** w-s < s
	Maximum time of glance duration	n.s.	** w-s > s	** w-s > s	n.s.
	Number of road-interface eye movements	n.s.	** w-s > s	** w-s > s	n.s.

The sonification barely influenced the lateral control of the vehicle. The only exception was a significant reduction of the SD of lateral position with auditory feedback, but only for task 1. On the other hand, longitudinal control was influenced by sound in different ways depending on the task: increased average acceleration in

task 1, decreased SD of acceleration in task 3, decreased SD of acceleration and increased inter-vehicular time in task 4. Although this pattern of result is not entirely consistant, it suggests a moderate facilitation of vehicular control with sonification.

Table 1 shows that the presence of sounds has a significant impact on the eye-tracking variables, mainly for the longer tasks (task 2 and 3: manipulation of GPS in a curve and in a straight line). In these cases, the presence of sounds significantly decreased the glance duration, the glance rate, the maximum time of glance duration and the number of eye movements toward the interface. The subjects spent less time looking at the interface with sounds than without, increasing driving safety. This effect was not observed for the shorter tasks (tasks 1 and 4 – reaction to a pop-up message). For task 4 (pop-up in a curve), the glance rate on the interface was even higher with sounds than without. For these two short tasks, to study the reactions of the driver after the task completion (not considered in the previous DVs), we looked at the number of visual controls of the driver on the interface after the task completion (controls to verify that the pop-up actually disappeared). The proportion of visual controls for each task and each condition is given in table 2.

Table 2. Proportion of visual controls on the interface after completion of tasks 1 and 4 in the two sonification conditions (without sound: w-s – with sound: s). Significance test of the difference (unilateral test of two proportions z-test - p<.05 - **p<.01 – n.s. : not significant (p>.05))*

	Task 1		*Task 4*	
	w-s	s	w-s	s
Proportion of visual control after the completion of the task	21/96	15/96	17/96	5/96
p-value (unilateral z-test of two proportions)	n.s. (p>.17)		** (p<.01)	

The results show that for task 4, the presence of sounds significantly decreased the proportion of visual controls after the task completion, and hence increases the driving safety.

Questionnaire

The questionnaire reports the subjects' subjective assessments, stated just after the experiment. The analysis of the response can be used to give indications to better understand the causes of the significant effects of the sonification condition on the DVs, based on the feelings of the drivers. For the questions about feeling of safety and feeling of not looking at the road, the responses rates are given in figure 6 and 7.

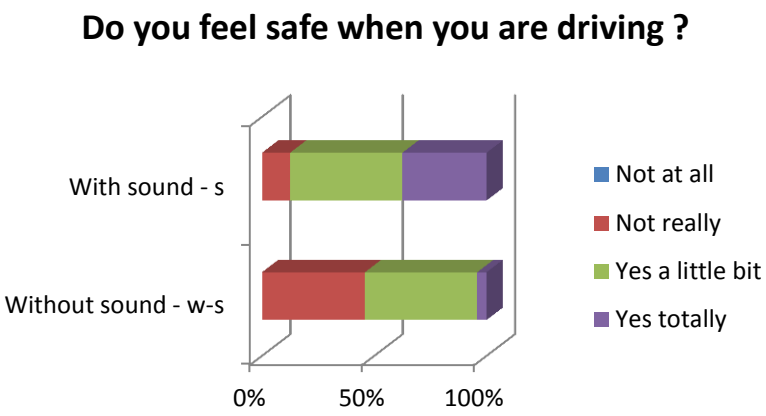


Figure 6. Proportion or responses to the question about safety

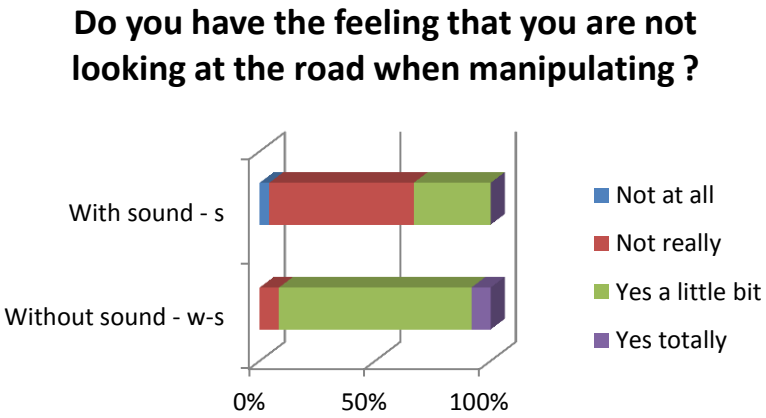


Figure 7. Proportion or responses to the question about "looking at the road"

The results show that the proportions of responses between the two conditions are significantly differents for the two questions (multinomial goodness of fit test, $p<.01$). With the presence of sounds, the subjects judged the interface as easier to use and had the feeling of looking more at the road.
Unsurprisingly, 100% of participants preferred to use the interface with sound than without sound. This supports the idea that the sound is a useful parameter for the manipulation of the interface. Finally, concerning the quality of sound design, the participants considered that the sounds were not troublesome and 91% appreciated the sound.

Discussion

Auditory feedback for touchless interaction marginally influenced driving behaviour, mainly in the longitudinal control of the vehicle. One reason may be that the subjects were very much engaged in the driving task during the whole experiment, with or without sounds. Inter-individual differences between subjects were large, possibly due to differences in driving style and experience. The influence of the sound in that context was relatively weak. Furthermore, for the short popups tasks (tasks 1 and 4), it was difficult to observe possible changes in behaviour in the 1 or 2 seconds the tasks required to be completed. In particular, before the validation, the sound did not bring any information. It was only useful for the confirmation of the task completion.

On the other hand, the interest of auditory feedback for the manipulation of the touchless interface was clearly evidenced by the gaze analysis. Indeed, the sound influenced gaze behaviour in a very significant way. With the sound, the participants looked much less at the interface when navigating in the menus or to verify that a call/message was refused or accepted. Furthermore, the lower number of eye movements between the interface and the road and the shorter glance duration in the "sound" condition shows that the users are feeling much less confident in their navigation. They could consequently pay more attention to the road and the driving task.

The results of the questionnaire confirm this analysis: with sounds, users are feeling safer, thanks to a sound design that meets the requirements (easily interpretable by user, bringing information needed to the navigation and relieving the visual workload).

Conclusion

The study showed that the presence of auditory feedbacks for the manipulation of a touchless interface of an infotainment system in a car significantly decreased the eyes-on-the-interface time. The sonification provides useful information on the manipulation of the interface, information that can only be obtained through vision in the absence of sound. It may increase safety because the driver can be more attentive to the road. The interface seems also easier to use with sounds, and the sound allows a more user-friendly experience.

References

Denjean, S., Roussarie, V., Ystad, S., & Kronland-Martinet, R. (2013). An innovative method for the sonification of quiet cars. *The Journal of the Acoustical Society of America*, *134*, 3979.

Hermann T., Hunt A., & Neuhoff J.G. (2011). *The sonification Handbook*. Logos Publishing House, Berlin, Germany. ISBN 978-3-8325-2819-5.

Jakus, G., Dicke, C., & Sodnik, J. (2015). A user study of auditory, head-up and multi-modal displays in vehicles. *Applied Ergonomics*, *46*, 184-192.

Kajastila, R., & Lokki, T. (2013). Eyes-free interaction with free-hand gestures and auditory menus. *International Journal of Human-Computer Studies*, *71*, 627-640.

Tardieu, J., Misdariis, N., Langlois, S., Gaillard, P., & Lemercier, C. (2015). Sonification of in-vehicle interface reduces gaze movements under dual-task condition. *Applied Ergonomics*, *50*, 41-49.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for Sensory and Consumer Science*. John Wiley & Sons Ltd. ISBN: 978-0-470-51821-2.

Sodnik, J., Dicke, C., Tomazic, S., & Billinghurst, M. (2008). A user study of auditory versus visual interfaces for use while driving. *International Journal of Human-Computer Studies*, *66*, 318-332.

Wilcox, R.R. (2014). *Winsorized Robust Measures*. Wiley StatsRef: Statistics Reference Online.

Car-following techniques: reconsidering the role of the human factor

Antonio Lucas-Alba[1], María T. Blanch[1], Teresa Bellés[1], Ana M. Ferruz[1], Ana Hernando[1], Óscar M. Melchor[2], Luis C. Delgado[3], Francisco Ruíz[4], & Mariano Chóliz[5]
[1]Universidad de Zaragoza ,[2]Impactware, [3]Universidad de Granada, [4]Universidad Konrad-Lorenz, [5]Universitat de València
[1,2,3,5]Spain, [4]Colombia

Abstract

Keeping correct distance between vehicles is a fundamental tenet in road traffic. New road signs and markings appearing on motorways aid drivers in determining this distance. However, the 'Nagoya experiment' (Sugiyama et al., 2008) revealed correct distance made following safe while also eventually destabilizing traffic flow. When traffic becomes dense, most drivers keep the minimum safety distance and brake when the vehicle ahead decelerates. The resultant chain reaction along the entire line of closely following vehicles causes for no apparent reason a traffic stoppage, known as a 'phantom' or 'shockwave' jam. The car-following models of Sugiyama et al. found certain speeds, traffic densities, and inter-vehicular distances combined to congest traffic. Drawing upon these and other phenomena (e.g., wave movement in Nature), car following by Driving to keep Inertia (DI) was conceived by us as an alternative to Driving to keep Distance (DD). Three studies explored possible prevention of 'phantom' jams by adopting DI. Using a driving simulator, affective and behavioural measures were taken (N=113). The results comparing the efficiency of DI vs. DD are summarized. DI promoted a more stable driver trajectory, in cognitive-affective and behavioural terms, and lowered fuel consumption by about 20%.

Background

This paper compares the efficiency of two elementary car-following (CF) techniques. Traffic flow efficiency may be judged by the prevalence of four driving modes: acceleration, deceleration, idling, and cruising (Tong et al., 2000). Efficient traffic cruises; congested traffic speeds up and slows down, polluting, wasting time and money, exasperating drivers, and risking life. As developed nations adopted stricter road safety standards, road salubrity worsened. Vehicle emissions now claim as many lives as crashes do, and possibly more (Caiazzo et al., 2013).

CF models were first developed in the early 1950s. Two main modelling efforts since then are the *Newtonian* or engineering CF models and the human factor

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

models (Saifuzzaman & Zheng, 2014). The rationale behind engineering CF models is the possibility to appraise and formalise how drivers *naturally* follow each other. Characterising and parameterising Normative Driving Behaviour (NDB) have become important goals since the late 1990s (Brackstone & McDonald, 1999). Hence, human drivers' collective movement is observed in the context of how animals move in Nature, and then it is modellised and predicted. But rather than being a Nature issue, CF is *nurtured* by official criteria derived from such technical documents as the *Highway Capacity Manual* (TRB, 2010). Perhaps drivers practice certain NBD, but they also heed official advice: keep safety distance.

This advice stems from the engineering and human rationale shaping such historical programmes as the USA's Federal-Aid Highway Act of 1956 (Weingroff, 1996). During the 1920s to 1940s, soaring car ownership brought wealth and also fatalities and traffic jams. Authorities then had to base growth of an adequate motorway network on certain calculations. If 50,000 drivers go from city A to city B daily at a reasonable pace (say, 100 km/h), what road geometry and capacity (e.g., number of lanes) would be needed? The answer is straightforward: consider a standard car speed and braking time (taking gravitational force, and a standard friction coefficient). Then consider time needed to slow down from, e.g., the maximum official speed if a car ahead brakes suddenly. Traffic safely cruising through a given road section should result. The desired following distance, say, 2 seconds (s), is thereby set – shaped top-down. Drivers, however, normally flout limits. In England, 95.8% keep less than 2 s and 47.9% less than 1 s (Brackstone et al., 2002).

Talking about *road capacity* may be misleading. Topologically speaking, a bucket has a limited capacity and a hose (road) does not. What prevents roads from being functional is the way flows are *ordered*. Hence, congested roads express lack of road capacity beyond reason, but so pervasively that they have earned a metaphysical label: *phantom traffic jam* (Gazis & Herman, 1992). But, why should stoppages arise not due to a bottleneck (e.g., caused by lane loss)? To answer, a shift from modelling coupled vehicles is needed; now 'traffic flow is investigated as a dynamical phenomenon of a many-particle system' (Sugiyama et al., 2008; p. 2). The Nagoya experiment aimed to create an artificial traffic jam. Drivers followed each other in a circle whose perimeter was 230 m. Participants were instructed only: *follow the vehicle ahead in safety in addition to trying to maintain cruising velocity*. And so they drove and kept free flow. But when the number of drivers was increased to 22, fluctuations tripping backward easily broke the free flow. Eventually several vehicles had to stop for a moment to avert crashing.

At stake here is longitudinal mechanical waves (Cromer, 1981). *Keep safety distance* is good advice for coupling vehicles on a road section, but, when more than two cars follow, cars platoon into a nearly perfect medium for wave transmission. As shown by Sugiyama et al., at some point the oscillatory nature of flowing cars spread, backward, to form a soliton of 25 km/h. Cars platooned so nicely that drivers, by virtue of the instruction *follow the vehicle ahead in safety*, could not avoid propagating the corresponding disturbances. It did not matter if tight couplings and platoons came from external reduction of space (adding cars to the circuit) or from voluntary decision (leaving less than 1 s distance to the car ahead).

Considering wave mechanics, we either eliminate disturbances or deal with the medium transmitting them – the car-following platoon. The former are difficult to control, but not the latter. To cope with a lead oscillatory car (the shockwave origin), a following car must become shockwave proof. This remedy may be sought by reversing the goal of Sugiyama et al.: instead of observing the cause of congestion, seeking a means of prevention. To this end, two driving techniques (DD/DI) are compared to see if one is more effective, in cognitive-affective and behavioural terms, in promoting steadier travel. DD is Driving to keep Distance (from the lead car) and DI is Driving to keep Inertia (an adaptive, uniform speed) while car following. Proposing these two orthogonal driving techniques (aim for uniform distance vs. uniform speed) opposes the idea of NDB as a unique driving mode (Brackstone & McDonald, 1999) and assumes drivers can learn to follow a lead vehicle proactively by changing from an automatic to a controlled operative mode (Charlton & Starkey, 2011) and applying DD or DI as appropriate.

Overview of the studies

Goals

All three studies aimed to check if: A) the same driver could drive in DD and DI modes when following a lead 'disturbing' car; B) drivers could follow the driving techniques by heeding a 10 s instruction (three sentences); C) DD vs. DI differences in cognitive-affective and behavioural terms were significant (Blanch, 2015). The relevance of such emotions as anger, fear or anxiety in troubled CF contexts like congestion have been documented (Shinar & Compton, 2004; Zhang & Chan, 2014). Additionally, Study 3 (Ferruz, 2015) monitored the space occupied by eight virtual automaton DD drivers following either a DD or a DI participant.

Participants

All participants were licensed drivers (table 1). Some were students participating in exchange for academic credit; others were invited via billboards at nearby shops, driving schools, restaurants, and the like.

Table 1. Main demographics of participants

	Study 1 (Blanch, 2015)	*Study 2 (Blanch, 2015)*	*Study 3 (Ferruz, 2015)*
N	44	44	25
Gender	20 men/24 women	7 men/37 women	13 men/12 women
Age	23.3 years	20.7 years	21.3 years
Education	84.1% university	68.2% university	100% university
Driving experience	4.07 years	2.81 years	2.68 years
Km per year (%)	59.1% < 10,000	59.6% <10,000	44.0% <10,000

Design

The three studies shared the same experimental design, a repeated measures model controlling for order. Manipulation of driving technique (DD, DI) was the within-

subject factor. Order (DD-DI, DI-DD), randomly assigned, was the between-subjects factor. The set of dependent measures concerned cognitive-emotional and behavioural indicators (table 2). The participants' basic task consisted of advancing in a straight line, for 4 minutes on a simulated road, and following a vehicle accelerating and decelerating (until stopping) cyclically, similar to what occurs in very congested traffic.

Materials

The studies were conducted in two rooms at the faculty laboratories of a Spanish university: a booth where participants executed the tasks and an adjoining room with two-way glass and a monitor displaying the participants' psychophysiological responses. One main study objective was characterizing the psychophysiological activity under DD and DI. Skin conductance response (SCR) was recorded with an MP36 unit (BIOPAC Systems, Inc., Goleta, CA, USA) at a sampling rate of 50 Hz by using two disposable Ag-AgCl electrodes attached to the left hypothenar eminence. Mean SCR was calculated in microsiemens (μS) for all three experiments. The MP36 unit connected to a standard PC running Windows XP.

Self-report measures of affective state were also collected via the Self-Assessment Manikin (SAM), a nonverbal pictorial rating technique (Lang, 1980). SAM was applied to measure the affective state after task execution in the simulator. It provides data on three general affective dimensions: valence, arousal, and dominance. SAM has been widely used and validated in psychophysiological research and has normative data adapted to the Spanish population (Moltó et al., 1999). The valence scale ranges from 1 (pleasure) to 9 (displeasure). The arousal scale ranges from 1 (exciting) to 9 (relaxing). The control scale ranges from 1 (low dominance) to 9 (high dominance).

One of the earliest goals of this research was designing a 3D driving simulator able to run on a standard PC in distant workplaces and laboratories. *ReactFollower* (Impactware, 2014), based on UNITY software, was developed and customized to change certain parameters (e.g., speed, frequency of stop-and-go cycles, etc.) externally, via XML. The focus was on materialising the possibility to study DD/DI against different oscillatory patterns of the lead vehicle. Participants were shown three scenarios, always in one lane on a straight road: A) participant drives alone on the road (always in a natural position on the driver's virtual side of the vehicle); B) participant drives behind another vehicle travelling at constant speed of 3 m/s (10.8 km/h); C) participant drives behind another vehicle traveling with constant stop-and-go cycles of a sinusoidal function built at a mean speed of 3 m/s (data is presented only from C). Participants could control acceleration/deceleration of their vehicle only by pressing 'up/down' arrows on a computer keyboard. When 'up' was pressed, it accelerated and maintained a constant speed when released. When 'down' was pressed, it decelerated. Acceleration/deceleration was in discrete increments: to accelerate or decelerate continually participants had to press the keys repeatedly. The simplest option (keyboard) was preferred to enable all participants to use the software with basic hardware equipment, and to level differences in expertise with video game keyboards. Finally, no direction changes were intended, just regulating speed-distance in a straight lane. The driving simulator worked on an HP

TouchSmart iq522es computer with a 23-inch screen, NVIDIA GeForce 9300m GS video card and 4 GB RAM, Intel Core 2 Duo Processor T6400 2.00 GHz, and Windows 7 operating system. A precision Apple USB keyboard (PCB DirectIN V2012) was used. The simulator collected, among others, variables for speed, distance to leader, and fuel consumption (a gross estimate obtained considering variations in speed per frame, see table 2).

Procedure

Scenarios A/B were designed as controls. In scenario C, participants were asked to follow the lead vehicle and adopt one of two driving techniques (DD or DI) though they never received an explicit verbal label for either. The group performing the task in DD-DI order received this instruction first for DD: 'In the simulated driving scenario that you will enter, you will see a vehicle ahead of you and it will not move at a constant speed. Sometimes it will go faster or slower. We ask you to travel behind that vehicle as closely as possible without risking a collision.' Following this, they used the simulator and then were given the SAM scales. Afterwards, the instruction for DI was provided: 'In the simulated driving scenario, you will see a vehicle ahead of you and it will not move at a constant speed. Sometimes it will go faster or slower. We ask you to travel smoothly behind the vehicle and maintain a constant speed, without letting the lead vehicle move too far away.' Participants in the supplementary condition (DI-DD) read the same texts in reverse order.

Overview of main results

Data were subjected to a repeated measure ANOVA having two levels of driving orientation (DD, DI). Table 2 presents the main results concerning SCR, SAM scales (valence, arousal, dominance), and performance indicators (speed, distance, fuel consumption) from the three studies. Skin conductance was systematically and significantly higher for DD vs. DI in all three studies (S-1, $p < .001$; S-2, $p < .001$; S-3, $p = .046$, $\eta_p^2 = .16$ to .37). Regarding SAM subscales, differences concerning valence were significant only in Study 2, with DI being judged as more pleasurable than DD ($p < .001$, $\eta_p^2 = .58$). Arousal was significantly higher for DD vs. DI in all three studies (S-1, $p = .004$; S-2, $p < .001$; S-3, $p < .001$, $\eta_p^2 = .18$ to .49). Dominance was higher for DI in S-1 ($p < .001$, $\eta_p^2 = .27$) and S-2 ($p < .001$, $\eta_p^2 = .37$), but not in S-3 ($p = .11$). Regarding performance indicators: Average speed was lower for DI in all three studies (S-1, $p < .001$; S-2, $p < .001$; S-3, $p = .004$, $\eta_p^2 = .26$ to .35), and also speed variability (S-1, $p < .001$; S-2, $p < .001$; S-3, $p < .001$, $\eta_p^2 = .68$ to .85). Conversely, average distance to leader was always smaller under DD (S-1, $p < .001$; S-2, $p < .001$; S-3, $p < .001$, $\eta_p^2 = .57$ to .60). Finally, fuel expenditure was lower under DI in the three studies (S-1, $p < .001$; S-2, $p < .001$; S-3, $p < .001$, $\eta_p^2 = .75$ to .89).

Table 2. Means corresponding to main variables

	Study 1		*Study 2*		*Study 3*	
	DD	DI	DD	DI	DD	DI
Skin conductance	8.04	6.55	9.47	8.18	11.11	9.26
Valence	3.45	3.45	5.79	2.93	3.48	3.52
Arousal	3.93	5.07	3.11	5.61	4.24	5.76
Dominance	6.25	7.20	4.91	6.77	5.68	6.44
Speed (m/s)	3.08	3.05	3.07	3.03	3.07	3.03
Speed variability (m/s)	2.57	1.44	2.54	1.44	2.24	.99
Distance to leader (m)	6.60	11.90	7.70	17.60	9.25	19.40
Fuel expenditure (l)	19.4	15.0	18.6	15.1	19.7	13.9

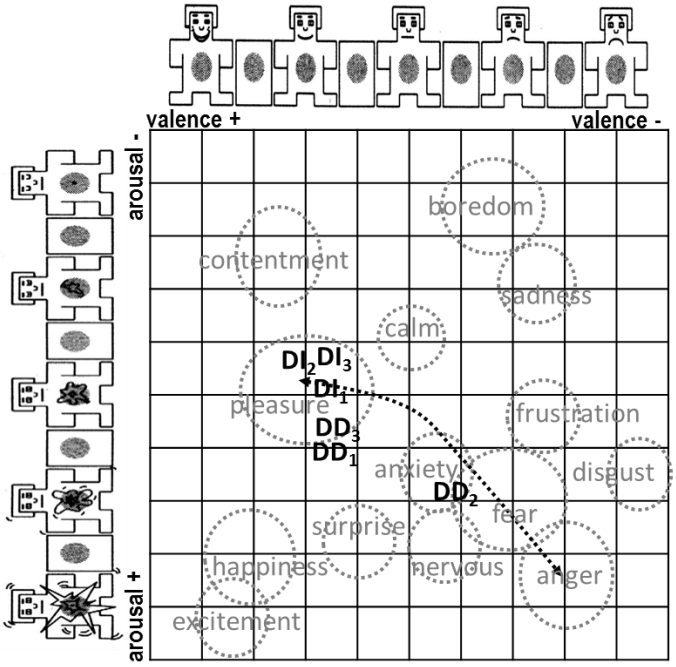


Figure 1. Mapping valence and arousal dimensions upon discrete emotions (Studies 1-3).

In sum, cognitive-affective indicators portrayed DI as a more comfortable way of following a lead oscillatory vehicle. SCR and SAM reports indicate DD drivers feel more arousal than DI ones (S1-3) and less dominance (S1-2), but only S2 shows valence differing. Following Cai & Lin (2011; see also Zhang & Chan, 2014), Fig. 1 tentatively maps results for valence-arousal dimensions (SAM) and discrete emotions. Performance indicators pointed to two orthogonal driving approaches, DD (aiming for uniform and shorter distance) vs. DI (aiming for uniform speed and longer distance). DI participants absorbed leader disturbance; moving at a more

uniform speed, they were in turn easier to follow (table 2). DD drivers kept a more regular, shorter distance to the lead vehicle, thereby sacrificing speed stability. DI drivers kept speed more uniformly, but needed more distance to cushion the lead car's stop-and-go pattern.

Results concerning the platoon of eight virtual following drivers

Study 3 included new measures by the simulator: eight virtual DD cars followed each participant, who was unaware of it. The simulator registered the distances between eighth vehicle and lead vehicle, and between eighth vehicle and participant. Average distance between eighth vehicle and lead vehicle was similar for each condition (DD: M = 117.3 m; DI: M = 118.95 m; p = n.s.). However, the distance between participant and leader was longer under DI (table 2), this fact obscuring the actual space required by the platoon. But differences between the eighth vehicle and the participant (DD: M = 108.03 m; DI: M = 99.55) were significant ($p < .001$; η_p^2 = .84). As measures of speed variability suggest (table 2), DI furnished platoon stability, and therefore optimised space on the road.

Discussion

DI drivers feel more comfortable, drive more steadily, and are easier to follow (even for DD virtual drivers). First, similar to differences found between car and truck drivers, the latter normally holding speeds more constant than the former (Ossen & Hoogendoorn, 2011), drivers in these three studies can drive under DD vs. DI mode when following a lead 'disturbing' car. Second, drivers can follow the driving techniques by heeding a 10 s instruction (three sentences) or a short video. Third, DI promotes a more stable driver trajectory than DD does, in cognitive-affective and behavioural terms. Fourth, all studies showed significant differences, always the same type, in these terms dependent upon whether participants applied DD or DI.

Potential relevance of training to learning DD/DI

Participants in the three studies received the same main instructions about the driving techniques. But compared with Studies 1 and 3 (short sentences described in *Procedure*), the set of instructions in Study 2 explained how to drive DD or DI with one of two videos (each 4 minutes approx.). Each video presented an explanation of congestion by one of two fictitious traffic institutes (named by acronyms, I.T.F.; C.M.D.). Both videos shared the same explanation for congestion (how congestion emerges), and then advised one of two behavioural alternatives (DD or DI). The main recommendation on how to drive was embedded (written) at the videos' end. Also, instruction for Studies 1 and 3 was direct, even more so than for Study 2 (Blanch, 2015). The difference in valence (SAM) in Study 1 and 3 vs. Study 2 is likely due to perceived authority of an agency (I.T.F.) recommending DD, the more stressful and harder to manage alternative (resulting also in higher arousal and lower dominance).

Limitations of the studies

This set of exploratory studies of DD/DI techniques contains some limitations. Compared with the average national driving population, study participants were more educated, younger, and unlikely to have driving habits ingrained by many years behind the wheel. Most were 'low mileage' drivers. They may have learned faster and been more amenable to new techniques than the average driver would be. Also, future studies should improve the ecological validity of *ReactFollower* (e.g., by using accelerator and brake pedals).

The main challenge, however, concerns comprehending how drivers' emotions, CF and congestion are linked. CF epitomizes the two elementary driving goals: safe/arrive. Inadequate distance concerns safety while slow speeds delay arrival. The literature shows anger is likely when drivers' goals are blocked by other drivers, and anxiety/fear emerge when drivers face probable danger (Mesken et al., 2007; Roidl et al., 2013; Zhang & Chan, 2014). Emotions, acting as a feedback loop concerning course of action, reset priorities and actions (Carver & Scheier, 2012). Congested CF increases opportunities for anger/aggression (tailgating, blocking of lane change for reaching exit), anxiety/fear (near rear-end crash) and relief (crash avoidance). This mix of emotions – Fig. 1's dotted line – may well cause oscillations in speeds and flow density. The data presented revealed differences in CF when either DD or DI was prompted, with an impact on arousal, but mobility goals – a key element concerning valence – were not manipulated. Future studies should analyse how emergence of certain emotions during DD/DI impact CF and congestion.

Concluding remarks

This paper aims to connect research on current car-following trends (Sugiyama et al., 2008) with operationalisation of two alternative driving techniques. For different reasons, drivers couple in dense traffic when lead vehicles are dictating the pace and keep a close, constant distance to each other. Learning a complementary way for adapting speed to oscillatory patterns of lead cars can contribute to alleviating congestion and its attendant ills while also stabilising successive car platoons.

Acknowledgements

We thank M. Pronin (A-Mazing Designs, NY) for improving the manuscript. We thank Alberto Arbaiza (DGT) and Professors Anxo Sánchez (Universidad Carlos III de Madrid) and José Luís Toca-Herrera (BOKU, Vienna) for assisting the research with insight and expertise. Support came from Fundación Universitaria Antonio Gargallo y Obra Social Ibercaja, Spain (grant 2015/B011).

References

Blanch, M.T. (2015). *El seguimiento de un vehículo en el desplazamiento en línea: caracterización psicofisiológica y conductual de dos formas básicas de conducción*. PhD thesis. University of Valencia, Spain.

Brackstone, M. & McDonald, M. (1999). Car-following: a historical review. *Transportation Research Part F, 2*, 181-196.

Brackstone, M., Sultan, B., & McDonald, M. (2002). Motorway driver behaviour: studies on car following. *Transportation Research Part F*, *5*, 31-46.

Carver, C.S. & Scheier, M.F. (2012). Cybernetic Control Processes and the Self-Regulation of Behavior. In R.M. Ryan (Ed.), *The Oxford Handbook of Human Motivation* (pp. 28-42). New York: Oxford University Press.

Cai, H. & Lin, Y. (2011). Modeling of operators' emotion and task performance in a virtual driving environment. *International Journal of Human-Computer Studies*, *69*, 571-586.

Caiazzo, F., Ashok, A., Waitz, I.A., Yim, S.H.L., & Barrett, S.R.H. (2013). Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005. *Atmospheric Environment*, *79*, 198-208.

Charlton, S.G. & Starkey, N.J. (2011). Driving without awareness: The effects of practice and automaticity on attention and driving. *Transportation Research Part F*, *14*, 456–471.

Cromer, A.H. (1981). *Física para las ciencias de la vida*. Barcelona: Editorial Reverté.

Ferruz, A.M. (2015). *Análisis de la intervención del factor humano en el movimiento vehicular en línea a partir de dos modelos de conducción*. Final Degree Thesis. University of Zaragoza, Spain.

Gazis, D.C. & Herman, R. (1992). The Moving and 'Phantom' Bottlenecks. *Transportation Science*, *26*, 223-229.

Lang, P.J. (1980). Behavioral treatment and bio-behavioural assessment: computer applications. In J.B. Sidowski, J.H. Johnson, and T.A. Williams (Eds.), *Technology in Mental Health Care Deliver Systems* (pp. 119-137). Norwood, NJ: Ablex.

Mesken, J., Hagenzieker, M.P., Rothengatter, T., & de Waard, D. (2007). Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports, (observed) behaviour, and physiology. *Transportation Research Part F*, *10*, 458-475.

Moltó, J., Montañés, S., Poy, R., Segarra, P., Pastor, M.C., Tormo, M.P., Ramírez, I., Hernández, M.A., Sánchez, M., Fernández, M.C., & Vila, J. (1999). Un nuevo método para el estudio experimental de las emociones: El International Affective Picture System (IAPS). Adaptación española. *Revista de Psicología General y Aplicada*, *52*, 55-87.

Ossen, S. & Hoogendoorn, S. (2011). Heterogeneity In Car-Following Behavior: Theory And Empirics. *Transportation Research Part C*, *19*, 182-195.

Roidl, E., Frehse, B., Oehl, M., & Höger, R. (2013). The emotional spectrum in traffic situations: Results of two online-studies. *Transportation Research Part F*, *18*, 168-188.

Saifuzzaman, M. & Zheng, Z. (2014). Incorporating human-factors in car-following models: A review of recent developments and research needs. *Transportation Research Part F*, *48*, 379-403.

Shinar, D. & Compton, R. (2004). Aggressive driving: an observational study of driver, vehicle, and situational variables. *Accident Analysis & Prevention*, *36*, 429-437.

Sugiyama, Y., Fukui, M., Kikuchi, M., Hasebe, K., Nakayama, A., Nishinari, K., Tadaki, S., & Yukawa, S. (2008). Traffic jams without bottlenecks –

experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, *10*, 3001-3007.

Tong, H.Y., Hung, W.T., & Cheung, C.S. (2000). On-Road Motor Vehicle Emissions and Fuel Consumption in Urban Driving Conditions. *Journal of the Air & Waste Management Association*, *50*, 543-554.

Transportation Research Board [TRB] (2010). *Highway Capacity Manual*. 5th Ed. National Research Council (USA). Washington, DC.

Weingroff, R.F. (1996). Federal-Aid Highway Act of 1956: Creating the Interstate System. *Public Roads*, *60*(1). Retrieved 25 January 2015: http://www.fhwa.dot.gov/publications/publicroads/96summer/p96su10.cfm

Zhang, T. & Chan, A.H.S. (2014). How appraisals shape driver emotions: A study from discrete and dimensional emotion perspectives. *Transportation Research Part F*, *27*, 112-123.

Comparing different types of the track side view in high speed train driving

Niels Brandenburger, Mareike Stamer, & Anja Naumann
German Aerospace Center (DLR e.V.)
Germany

Abstract

The introduction of high speed trains featuring an increasing number of automated components raises imperative questions concerning the future tasks and the general role of the train driver. Previous work showed that train protection systems provoke train drivers to relocate their visual attention from the track side towards the displays within the cabin. The introduction of high speed routes allowing automatic train operation (ATO) has major implications that question the importance of the track side view for the train driver: (1) all relevant driving parameters are displayed within the cabin in high speed railway operations. (2) Supervisory tasks based on in-cab display information shift into the train driver´s focus. This study investigated the influence of three differently sized track side views (real size, monitor size, none) on a) the allocation of visual attention towards displays and the track, measured by eyetracking parameters and b) the situation awareness of the train driver supervising a high speed train featuring ATO measured with the SPAM method. Empirical data are presented for both research questions. The implications are discussed in order to identify how the delivery of relevant information in the context of the changing train driver's task can be facilitated.

Introduction

Driving long distance trains is a rather monotonous job (Dunn & Williamson, 2011; Stein & Naumann, 2016), especially with a limited number of stops along the route as in high speed passenger transport.

We are continuously looking at increasing the efficiency of track utilisation and energy consumption as well as punctuality (European Commission, 2011) and after decades of technical development and debate automating high speed passenger operations is today seen as one valid way to achieve these efficiency gains. At the moment automatic train operation (ATO) is restricted to either urban transport operated as closed systems or certain track segments of low complexity. To ensure efficiency gains throughout the whole railway industry, ATO is likely to be introduced to more complex segments containing e.g. railroad switches, platforms, level crossings. As high speed tracks can be considered to be of moderate complexity because these infrastructure elements are less frequent, we choose to start our investigation on the train driver´s tasks in this operational context.

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Additionally, the high speed operational context is most suitable for our research questions, because in-cabin signalling already is the norm today, in contrast to e.g. freight train operations where train drivers still heavily rely on real world cues found along the track to verify their driving behaviour (Rose & Bearman, 2012). The key question obviously is where the main sources of relevant information are located within the working environment of the train driver and whether the drivers allocate their attention to these areas.

Previous work shows a shift of visual attention onto the displays within the cabin away from the track side view once a train protection system is at work (Naumann, Wörle, & Dietsch, 2016). These results are based on the train protection system PZB 90 (for a description see Naumann, Grippenkoven, & Lemmer, 2016), a German train protection system mainly relying on signal aspects and speed indicators located along the track, while the display offers information concerning the status of the train protection system. Therefore the shift of visual attention in the PZB 90 condition compared to a condition without train protection in a simulation study described by Naumann, Wörle, & Dietsch (2016) is in fact a shift away from the information source delivering the major parameters for deriving driver instructions. Even more so because train drivers driving under PZB 90 supervision do have the explicit order to monitor the track ahead for obstacle detection. In contrast, in current high speed passenger operations relevant information is being displayed within the cabin so that is where the visual attention is supposed to focus on. Technically, visual obstacle detection is still a task of legal relevance to the high speed train driver as well, but due to breaking distances which exceed the train driver´s field of view by far, the need for technical obstacle detection and physically enclosing high speed tracks is undisputed. Even more so if ATO is to be implemented, thus emphasizing the critical need to monitor the driving parameters within the cabin, which are likely to be displayed using the ETCS - DMI (European Train Control System - Driver Machine Interface) display layout (European Railway Agency, 2007). This major shift of attentional resources into the cabin (Naweed, 2013) inevitably calls for continuous display monitoring as a key characteristic of automated high speed passenger operation out of two reasons. On the one hand the display information shows what kind of driving behaviour is required in accordance with the track environment and on the other hand the current modus operandi of the automated train components is displayed within the cabin as well. Therefore the continuous comparison between required and automatically executed driving behaviour is essentially based on display information. Previous research (Brandenburger & Naumann, 2016) has identified a lack of clarity in the understanding of the role the track side view is playing in such an operational scenario. Additionally, experienced train drivers claim that the track view is to stay a central part of the train driver´s job in ATO in order to integrate e.g. track conditions or geographic orientation into driving behaviour. Therefore, the first research question aimed to assess (a) to what extend the track side view out of the cabin windows is actually used under ATO conditions in contrast to a non-automated condition and whether the size of the track side view alters the extend of usage. Given the characteristics of the continuous display monitoring task to be executed by the train driver supervising an ATO, there is already a sound body of scientific knowledge applicable to the task at hand. Dunn and Williamson (2011) provide

insights on monotony as a consequence of repetitive task characteristics in railway passenger operations and Edkins and Pollock (1997) identify sustained attention and especially inattentiveness to railway signals as the major factor present in all types of railway accidents in a review of train accidents over a 3-year period. Spring et al. (2008) indicate poorer vigilance under ATO in comparison to manual driving with in-cabin signalling and over speed intervention. The authors also report an additional vigilance decrement only present in the ATO group over the course of their experiment. Warm, Parasuraman and Matthews (2008) also characterize cockpit monitoring as a vigilance task hinting at the differentiation between simultaneous and successive vigilance to clarify under which tasks vigilance decrements are more likely to be found. They proclaim successive vigilance tasks to be more vulnerable as the ”observers need to compare current input with a standard retained in working memory to separate critical signals from nonsignal stimulus events” (Warm, Parasuraman & Matthews, 2008; p.435) in contrast to simultaneous vigilance task where ”all the information needed to distinguish signals from nonsignals is present in the stimuli themselves” (Warm, Parasuraman & Matthews, 2008; p.435). While manual driving under in-cabin signalling is mainly relying on verifying manual traction input visualized by means of the speedometer needle against a visually presented speed limit representing a simultaneous vigilance task, monitoring and supervising ATO does include upholding a complex mental model of the automatic components resulting in a certain traction adjustment. Therefore, it can be argued that this task is more successive in nature. Kaber and Endsley (2004) argue that monitoring automated systems is associated with poor vigilance and a failure to build up and maintain an accurate mental model. This model contains an adequate understanding of the underlying automatic components, their functionalities and how these functionalities translate into real world in this specific case traction adjustment matching trackside train protection system´s requirements. Thus, is argued to result in loss of situation awareness (Kaber & Endsley, 2004). Building on a large body of empirical evidence it is undisputed that the perception of relevant information is the key to develop and maintain situation awareness, eventually anticipating future action of the monitored system (Endsley, 1988; Endsley, 1995; Parasuraman, Sheridan & Wickens, 2008). Therefore, the second research question arises (b) whether the size of the track side view (displaying irrelevant information) influences measures of situation awareness in ATO. If so, this has implications in terms of unnecessary distraction of the train driver, who´s central task is to monitor the in-cabin displays.

Hypotheses

Based on Naumann, Wörle, & Dietsch (2016) we hypothesised a shift of visual attention from the trackside view to the display (DMI) in a more automated driving condition compared to a manual driving condition.

> H1) *We expect the number of fixations on the DMI to be higher in the ATO condition than in the manual condition* (figure 1).

As the salience of the track side view is thought to decreases with smaller size

H2) *We expect the number of fixations on the DMI to increase with decreasing size of the track side view* (figure 1).

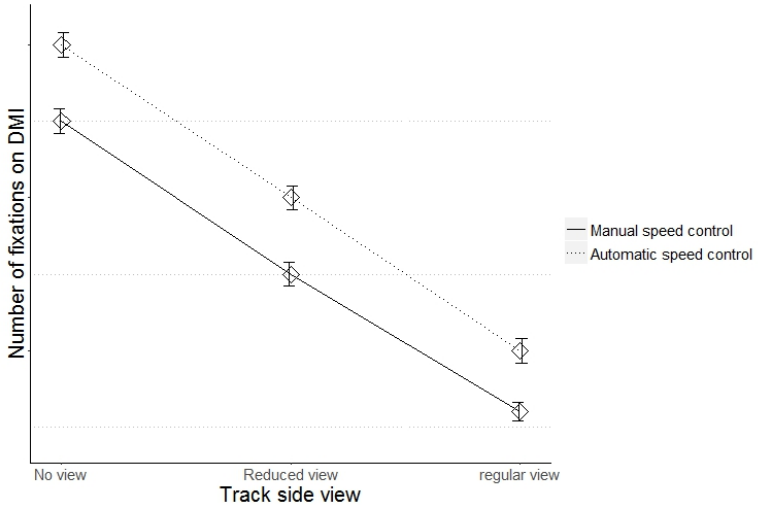


Figure 1. Hypothesized relationship of Number of fixations on DMI and size of Track side view for driving conditions

Based on the findings by Kaber and Endsley (2004) concerning level of automation and situation awareness

H3) *We expect the situation awareness measures to be smaller in the ATO condition in contrast to the manual condition.*

As the information contained in the track side view in our experimental setting representing future high speed track infrastructure is irrelevant to the monitoring task

H4) *We expect the situation awareness measures to increase with decreasing size of the track side view.*

Method

Participants

26 male German train drivers aged from 21 to 56 (M = 36.53, SD = 10.92) were recruited out of a pool of previous participants. Participants worked in freight (4 out of 26) and passenger transport (22 out of 26). Their occupational experience was ranging from 1 to 37 years (M = 14.07, SD =10.85). All were unfamiliar with the ETCS system featuring automatic speed control. Train drivers with glasses and train drivers following regular medical treatment were excluded from participation. The sample was randomly assigned to the experimental conditions and participants were compensated with € 30.

Experimental Design and Measures

Figure 2. High fidelity railway simulator while driving. In this case with regular sized track side view and the ETCS-DMI giving permission to drive up to 400 km/h.

The study was designed in a 2x3 mixed design including both the variable "speed control" with two between subject levels "manual speed control" / "automatic speed control" and the variable "track side view" with three within subject levels "regular view", "reduced view" and "no view". Participants drove in the high fidelity railway simulator "RailSET" (figure 2), which is presented and explained in detail by Stein and Naumann (2016). The simulator incorporates a cabin (BR 424 manufactured by Alstom/Siemens), a projector (SONY VPL-FH500L), two monitors as side windows (100 cm screen diagonal) and three monitors (30 cm screen diagonal) in the desk as instruments, from which only the center monitor was used during the experiment to display the ETCS-DMI (European Railway Agency, 2007). For experimental manipulation, the simulator was equipped with automatic speed control functionality and three ways of displaying the track side view. The regular track side view includes full frontal presentation of the 3D simulation environment and additional side windows. A reduced track side view consisted of a reduced frontal presentation and no side windows and in the third option no frontal or side view of the 3d simulation was present. The DMI was visible in all conditions. Furthermore, measurement of the dependent variables made the following apparatus and materials necessary. Eyetracking data was obtained using a head mounted Dikablis Essential system in combination with the software DLAB Version 3.0 both supplied by

Ergoneers. Subjective data was assessed using paper and pencil questionnaires of the following kinds. The Situation Awareness rating technique (SART) by Taylor (1990) was first translated to German and adapted to the high speed railway context and then used to assess subjective situation awareness. The adapted SART comprised 10 questions assessing three areas on a 4-point Likert scale (Demand from Attentional Resources (D), Supply of Attentional resources (S), Understanding of the Situation (U)). The SART score was the sum of U and S minus D (Taylor, 1990), with a theoretical range from -5 to 27. Additionally, the situation present assessment method (SPAM) proposed by Durso et al. (1998) was implemented using 12 questions targeting current and future understanding of the driving situation. The SPAM scores were the answering times for correct answers. Further paper and pencil materials were a demographic questionnaire, an informed consent, a short explanatory handout of ETCS and the underlying functionalities of the simulator and a debriefing form. The explanatory handout was used in two versions, one of which incorporated information about the automatic speed control function used in one of the experimental conditions. Lastly, a voice recorder was present to record the answers to the SPAM questions and for post hoc artefact detection.

Procedure

The study took place in the research facilities of the German Aerospace Center in Braunschweig, Germany. Upon arrival the participants gave their informed consent before filling in the demographic questionnaire and reading the information on the simulator. Then the eyetracking device was attached and calibrated. Afterwards, the participants started driving the first of three experimental blocks for approximately 35 minutes. During the driving block the experimenter shortly joint the participants two times in the cabin (after approximately 10 and 20 minutes of driving) to ask two questions at a time related to the SPAM method. The driving block ended with an ordered standstill of the train conveyed through the DMI. In the pause between the first and the second driving block the participant filled in the adapted SART questionnaire. The second and third driving block as well as the pause between them followed the routine described for the first driving block and the first pause. After finishing the third driving block the participants once more filled in the SART questionnaire. Afterwards all recording equipment was shut off and detached before participants got a debriefing along with their monetary compensation and were dismissed.

Results

Attention allocation

Based on findings of Naumann, Wörle, & Dietsch (2016), we formulated our first hypothesis that we expect the number of fixations on the DMI to be higher in the ATO condition compared to the manual driving condition. Nevertheless the between-subject effect for the driving condition was not found significant ($F(1,24) = .486$; $p > .05$) employing a repeated measures ANOVA including both factors track side view (three levels) and driving condition (two levels) as independent variables and the number of fixations on the DMI as a dependent variable. Therefore we could not reject the null hypothesis based on sample data. Inspecting figure 3

added additional evidence to the fact that the difference between the group means, although showing larger numbers of fixations for the ATO condition, is not of significant size in relation to the within group variations in both driving conditions, also reflected by a small to moderate effect size (η^2 =.20).

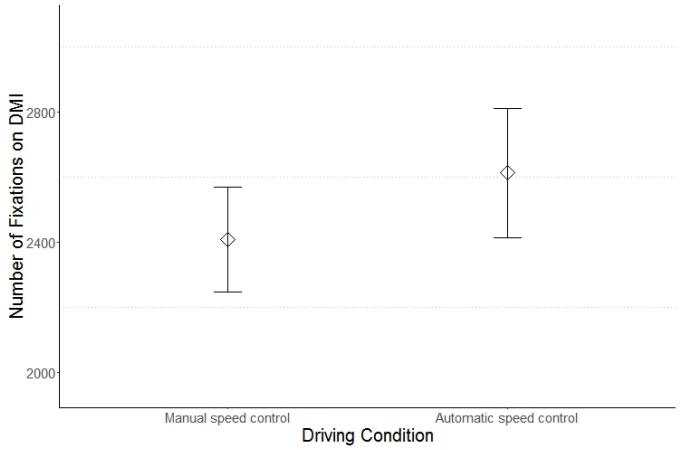


Figure 3. Absolute Number of Fixations on the Driver- Machine- Interface in our sample by driving conditions. Error bars represent the standard error of the mean (SE).

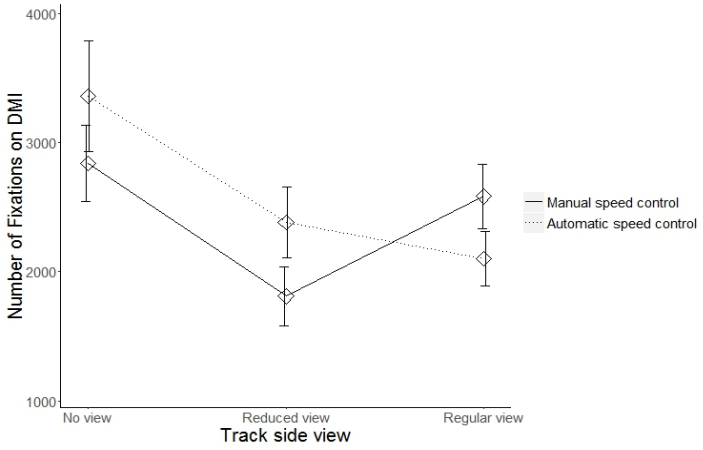


Figure 4. Absolute Number of Fixations on the Driver-Machine-Interface for three different sizes of Track side view by speed driving conditions. Error bars represent the standard error of the mean (SE).

Concerning the second hypothesis, a highly significant within-subject effect for the size of the track view on the number of visual fixations was found again using repeated measures ANOVA while relying on Greenhouse-Geyser corrected degrees of freedom (F (1.817, 43.613) = 8.86, p <.01). The size of this partial effect was

moderate (η^2 = .270). Accordingly, the data displayed in figure 4 indicated an increasing number of fixations on the DMI with decreasing size of the track side view, which is in line with hypothesis 2. Nevertheless, figure 4 showed an interesting deviation from this pattern in the regular track side view condition, where the number of fixations was actually higher for manual speed control condition, raising questions concerning a possible interaction effect. Nevertheless, there was only a trend for an interaction effect in a repeated measures ANOVA model ($F(1.817, 43.613) = 2.859$, $p = .073$) failing to reach significance. In the absence of a significant interaction effect figure 4 once more shows that indeed the number of fixations on the DMI steadily increased with smaller track side view irregardless of speed control. Although, the curve for the manual driving condition deviated somewhat from our expectations voiced in hypothesis 2 in the sense that regular size track side view resulted in more fixations of the DMI in the manual condition than expected, we accepted hypothesis 2 after having ruled out an interaction effect undermining main effect relevance.

Situation Awareness

To test hypotheses 3 and 4, both SART and SPAM scores were evaluated. Concerning hypothesis 3, claiming situation awareness ratings to be smaller in the ATO condition, a repeated measures ANOVA testing for differences in the dependent SART scores did not reveal a group mean difference between driving conditions ($F(1,19) = 1.104$, $p > .05$). Similarly, no between-group effect was found for the SPAM scores ($F(1,22) = .267$, $p > .05$), which is also visible in the small sizes of the effects for SART (η^2 = .055) and SPAM (η^2 = .012) measures. Interestingly, the Pearson correlation between the SART and the SPAM scores was small ($r = .340$). Therefore, hypothesis 3 was not validated.

The fourth hypothesis claimed that an increased situation awareness rating would go along with decreased size of the track side view. Nevertheless repeated measures ANOVAs for SART ($F(1.945, 36.964) = 3.553$, $p > .05$) and SPAM ($F(1.727, 37.989) = .679$, $p > .05$) did not show a significant within subject effect of the track side view factor. These results did not validate hypothesis 4. Even more so figure 5, depicting subjective SART scores for different track side views driving conditions, did show a trend opposite to hypothesis 4, showing slightly better situation awareness ratings for an increasing size of the track side view. Objective SPAM scores remained mostly unchanged over the track side view conditions (figure 6).

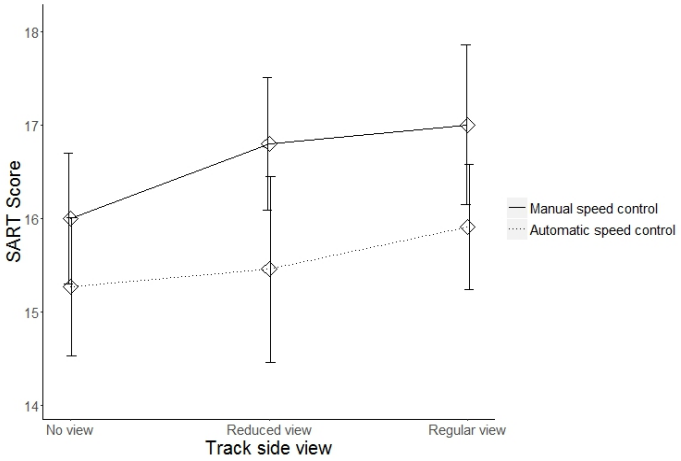


Figure 5. Subjective SART Score for three different sizes of Track side view by driving conditions. The depicted SART Scores theoretically range from -5 to 27. Error bars represent the standard error of the mean (SE)

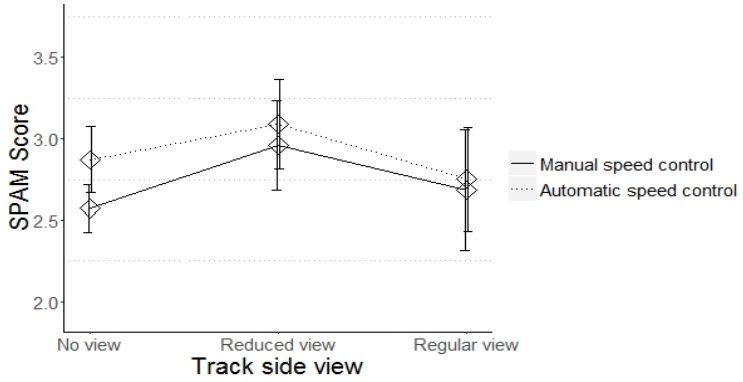


Figure 6. Objective SPAM Scores in seconds for three different sizes of Track side view by speed driving conditions.

Discussion and Conclusion

Aim of the current line of research is to identify how relevant information about a certain train ride can be delivered effectively to the train driver given the fact that ATO is at work. This does question central assumptions concerning the train driver´s tasks, two of which we stated in the introduction for further investigation. Namely, (1) all relevant driving parameters are displayed within the cabin in high speed railway operations and (2) supervisory tasks based on in-cab display information shift into the train driver´s focus, basically rendering track supervision

irrelevant in terms information acquisition of driving parameter. Therefore the study investigates a) whether ATO (hypothesis 1) and decreasing size of the track side view (hypothesis 2) containing irrelevant information lead to an increased focus of visual attention on the in-cabin DMI. Additionally it is investigated b) whether ATO (hypothesis 3) and decreasing size of the track side view (hypothesis 4) result in heightened situation awareness of the train driver. Concerning research question (a), a significantly higher number of fixations on the DMI for train drivers supervising ATO (hypothesis 1) cannot be found. Nevertheless, the observed group differences between the ATO and the manual driving group are in line with hypothesis 1 and findings from Naumann, Wörle, & Dietsch (2016), who reported train protection functionality to result in more visual attention on the DMI. As hypothesized, a decreased size of the track side view is found to result in more visual attention allocation on the DMI (hypothesis 2). This effect is especially prominent in the ATO condition resembling future high speed train operation. Closer inspection of the data (figure 4) revealed unexpectedly high numbers of fixations on the DMI for manually driving train drivers equipped with a regular outside view. Even in the absence of a significant interaction effect this deviation is worth noting as it contradicts the current understanding that modern in-cabin signalling along with automatic train protection forces the train driver´s attention onto the displays (Naumann, Wörle, & Dietsch, 2016; Naweed, 2013). It seems crucial to ensure that in-cabin signalling as e.g. in the ETCS- DMI is side-lined by measures enhancing the attention allocation onto the relevant displays, especially in ATO environments. Ultimately, results on hypothesis 2 lead to the conclusion that the size of the track side view may be employed to redirect visual attention towards in-cabin equipment in an ATO environment. The effects of track side view size on visual attention in manual driving environments need further clarification.

Concerning the train driver´s situation awareness (research question b) in ATO environments, several results are of interest. First of all, reduced situation awareness in the ATO condition as a negative consequence of the higher automation functionality for both of the situation awareness measures was not found (hypothesis 3). Moreover, objective situation awareness (SPAM) tends to be higher in the ATO condition in our data, while subjective situation awareness (SART) tends to be lower in the ATO condition. The correlation between the subjective SART measure and the objective SPAM measure was small in size. One possible explanation for this inconsistence is the moderating effect of time pressure described by Stoller (2013) who reports that correlations between subjective and objective situations awareness measures decrease if time pressure is low. The results concerning the effect of track side view size on situation awareness are statistically inconclusive (hypothesis 4), as the hypothesised effect failed to reach significance. Interestingly, subjective situation awareness tends to grow with size of track side view, while objective Situation Awareness remains constant. The presented results lead to the conclusion that varying the size of the track side view containing mainly irrelevant information to the train driver does not seem to endanger a proper level of situation awareness in automatic high speed passenger operation. Nevertheless, the size of the track side view as a unique source for information on orientation and weather conditions and its impact on situation awareness in current railway operations still needs further investigation.

Future research questions in this line of research will also focus on the effects of part-time supervision in contrast to full time supervision of automatic high speed trains. Likewise facilitating adequate responses to failures in automatic speed control components as well as implementing the insights into a working environment enabling high speed passenger train operation to be supervised remotely are of interest.

References

Brandenburger, N., & Naumann, A. (2016). Determining the allocation of tasks in automated high speed driving. In B. Milius & A. Naumann (Eds.), *Proceedings of the 2nd Workshop on Human Factors* (pp. 112-121). Braunschweig: ITS automotive nord.

Dunn, N., & Williamson, A. (2011). Monotony in the rail industry. In *HFESA Conference* (47th ed.).

Durso, F.T., Hackworth, C.A., Truitt, T.R., Crutchfield, J., Nikolic, D., & Manning, C.A. (1998). Situation Awareness As a Predictor of Performance for en Route Air Taffic Controllers. *Air Traffic Control Quarterly, 6*(1), 1-20.

Edkins, G.D., & Pollock, C.M. (1997). The influence of sustained attention on Railway accidents. *Accident Analysis & Prevention, 29*, 533-539. doi:10.1016/S0001-4575(97)00033-X

Endsley, M. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 Aerospace and Electronics Conference*.

Endsley, M. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors, 37*, 32–64.

European Commission (2011). Weissbuch: *Fahrplan zu einem einheitlichen europäischen Verkehrsraum – Hin zu einem wettbewerbsorientierten und ressourcenschonenden Verkehrssystem*. Brussels, Belgium.

European Railway Agency (2007). *ERTMS/ ETCS: Functional Requirements Specification FRS*.

Kaber, D.B., & Endsley, M.R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science, 5*, 113–153. doi:10.1080/1463922021000054335

Naumann, A., Grippenkoven, J., & Lemmer, K. (2016). Future information and assistance systems for train drivers and evaluation of their usability. In: *Proceedings of the 11th World Congress on Railway Research*, Milan, May 29 –June 2, 2016.

Naumann, A., Wörle, J., & Dietsch, S. (2016). The effect of train protection systems on train drivers' visual attention. Poster at the HFES Europe chapter annual meeting, 2016. http://www.hfes-europe.org/posters-2016/

Naweed, A. (2013). Investigations into the skills of modern and traditional train driving. *Applied Ergonomics, 45*, 462–470. doi:10.1016/j.apergo.2013.06.006

Parasuraman, R., Warm, J.S., & Dember, W.N. (1987). Vigilance: Taxonomy and utility. In L.S. Mark, J.S.Warm, & R.L. Huston (Eds.), *Ergonomics and human factors: Recent research* (pp. 11–32). New York: Springer-Verlag.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making, 2*, 140–160. doi:10.1518/155534308X284417

Rose, J., & Bearman, C. (2012). Making effective use of task analysis to identify human factors issues in new rail technology. *Applied Ergonomics*, 43, 614–624. doi:10.1016/j.apergo.2011.09.005

Spring, P., McIntosh, A., Caponecchia, C., & Baysari, M. (2008). Level of Automation: Effects on Train Driver Vigilance. *44th Annual Human factors and Ergonomics Society of Australia Conference,* 264–271.

Stein, J., & Naumann, A. (2016). Monotony, Fatigue and Microsleeps - train drivers` daily routine: a simulator study. In B. Milius & A. Naumann (Eds.), *Proceedings of the 2nd Workshop on Human Factors* (pp. 96–102). Braunschweig: ITS automotive nord.

Stoller, N. (2013). *Situation Awareness von Lokführenden während sicherheitskritischer Ereignisse im Bahnverkehr* (Bachelor Thesis). Fachhochschule Nordwestschweiz Olten.

Taylor, R.M. (1990). Situation awareness rating technique (SART): the development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (Chapter 3). France: Neuilly sur-Seine, NATO-AGARD-CP-478.

Warm, J., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors, 50*, 433–441.

A framework for human factors analysis of railway on-train data

Nora Balfe
Centre for Innovative Human Systems, Trinity College Dublin
Ireland

Railway operations are increasingly captured using digital technologies, such as on-train-data-recorders (OTDR) which record all control inputs by the train driver along with other metrics including speed, distance travelled, and GPS coordinates. Already widely used for fleet management and fault finding, the data may also have a potential human factors application in analysing and improving railway operations, for example by providing leading indicators of train driver performance or highlighting infrastructure sections correlated with poor driving performance across all drivers. This research explores the possible use of such data, and the barriers to be overcome in its application, including the size of the data sets, the reliability of the data and the identification of useful features or metrics within the data. A framework of a typical train journey is presented, breaking the journey into segments within which the OTDR data can be analysed. The key metrics available from the OTDR that may be applicable to each journey segment are discussed, along with the potential benefits of utilising the data in this context and a roadmap for future research in the area.

Introduction

Human factors research on the train driving task dates back as least as far as Branton, who in 1979 published a paper discussing the nature of train driving and the need for drivers to anticipate future actions, develop internal representations of the railway (called route knowledge), and test these representations against reality. Today, rail human factors is a growing and vibrant discipline, which examines the ways in which human performance in railway operations can be supported (Wilson & Norris, 2005). Changes in technology generate parallel changes in the role and demands of the human operators, many of which are positive but some of which have unexpected consequences and the human factors discipline is constantly developing methods and approaches to study, analyse and improve the system. This paper presents and discusses the possible contribution of on train data recorders (OTDR) in this context.

Digital train event recorders are increasingly placed on board rail vehicles, capturing information on train movements and component status, but currently this detailed information is not widely used outside of incident investigations and maintenance management (de Fabris et al., 2008; Mosimann & Rios, 2014). Walker and Strathie

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

(2014) suggest that train recorder data is an underused but potentially important data source for understanding human performance and detecting risks in advance of accidents. This is particularly important in the context of the current safety performance of the rail industry, characterised by very few major incidents and relatively stable safety performance indicators. There are no major, obvious changes to be made to improve safety and new ways of looking at and using data are needed to give further insight into issues and potential improvements, as in the Safety II approach (Hollnagel, 2014), which proposes that human performance variability is a key part of system resilience and by studying how this variability contributes to good system performance we can better manage safety.

As in other safety critical industries, human performance is a major contributor to safe railway operations, with an analysis by Evans (2011) finding that 69% of fatal railway accidents across seven European countries between 1980 and 2009 were primarily related to human performance. The train-driving task is primarily visual-spatial, involving constant perception and processing of information (Gillis, 2007) and the majority of train driver physical actions are driven by information received (e.g. moving the traction handle in response to a change in the speedometer), placing a strong requirement for visual attention on the train driver. Key tasks involve processing information collected from both inside and outside the cab and applying route knowledge to correctly control the speed and braking of the train (Doncaster, 2012; Hamilton & Clarke, 2005; Buksh et al., 2013). However, despite the apparent simplicity of the task, Naweed (2014) describes the train-driving task as complex, dynamic, and opaque. Although the basic tasks may be described reasonably simply, the actual practice involves changing conditions, event densities, and performance pressures that drive adjustments in motor skills and problem solving strategies. The complexity is driven by sometimes conflicting goals of time-accuracy, comfort, and speed regulation and the trade-offs required to optimise the overall journey. The dynamism comes from the constant need to regulate speed and finally, the opacity is due to the gaps in information when working with lineside signalling; drivers must use their route knowledge to infer future requirements. Thus, train driver performance is not simply a matter of perceiving and responding to stimuli as suggested by the use of simple information processing models, but is driven by continuous, proactive predication and planning (Elliott et al., 2007).

A key question in monitoring and managing driver performance is, what are the attributes of good train driving? There is very little information in the literature directly addressing this question, although some papers list the attributes of good drivers (e.g. strong mental models, ability to anticipate, good concentration; Russell & Long, 2005). Some experimental studies have used measures such as use of braking power, adherence to speed limits, variability of speed, and reaction to warnings as dependent variables linked to driver performance (Dorrian et al., 2005, 2006; Dunn & Williamson, 2012; Robinson et al., 2015) but there has not been specific validation of such measures as general indicators of driver performance, or any work to integrate different measures into a unified model. Driver performance measures are starting to be widely used in the road transport domain, where they can be collected by in-vehicle data recorders (IVDR) These devices can identify undesirable driving behaviours, such as speed, hard braking, accelerating, sharp

turning, and swift lane changes (Albert et al., 2011) and, in specifically equipped vehicles, even driver gaze and distance to lateral road marking (Pérez et al, 2010). In current implementations, this information is typically communicated to driver live while they are driving or as a performance summary at the end of the trip. Research has suggested that such feedback is linked to improved performance (e.g. Musicant et al., 2010; Donmez et al., 2008) and similar benefits may be attainable from the use of OTDR to provide feedback on train driver performance. Professional road drivers tend also to be monitored on energy efficiency, which may also be possible in the rail environment using OTDRs.

Research in OTDR use in railway performance monitoring is more limited than IVDR, but a small number of studies have been published. Two studies (Walker & Strathie, 2014; Rashidy et al., 2016) analyse OTDR data to explore train driver interaction with warning systems. Interesting findings include a high false alarm rate (Walker & Strathie, 2014) and a consistently speedy response to alarms from some drivers (Rashidy et al., 2016). Such research provides highly useful insights into the effectiveness of real-world warning devices and highlights areas for possible improvements, but only utilises a small portion of the available data. Strathie and Walker (2015) have also applied link analysis and associated graph theory to the analysis of on train data recorders. The analysis linked each control action by a driver to their next control action in a diagrammatic form, and facilitated the further analysis of driver styles and differences. The results found that there was a difference in the number of links between elements of the control interface (i.e. some drivers moved between more pairs of controls than other drivers). The number of links was found to be fairly consistent within some drivers, i.e. they reflected a stable driving style of that particular driver, but variable across others. However, the data cannot be used to determine whether this reflects an inherently unstable driving style or external factors motivating different behaviours on each journey analysed. Other analyses, such as the sociometric status of a node, the mean number of throttle moves per journey, the network diameter (or number of links in a chain of movements), all also showed promise for differentiating between driver styles. These indicators have not yet been linked to ‘good’ driver performance, but many ultimately provide a way to differentiate between good and poor driving behaviours.

Finally, Green et al. (2011) proposed the use of OTDR data in assessing driver performance. They note that such data is used in a qualitative fashion in current competence management, but that there is very little attempt to classify drivers according to their driving ability or risk. They identified five routine events for an initial implementation of driver performance monitoring, but did not provide any justification or evidence supporting these as valid metrics able to distinguish between different levels of driver performance:

1. Speed at which power notch 4 is selected when accelerating (passenger comfort)
2. Percentage of time in a braking sequence that the driver selects brake step 3 (passenger comfort)
3. Speed over TPWS grids approaching a Permanent Speed Restriction (PSR) (train speed)

4. Speed through a PSR as a percentage of the maximum speed (train speed)
5. Mean speed when an AWS horn is received (train speed)

These metrics can all be automatically calculated from the OTDR downloads, and displayed on a dashboard showing each drivers' performance. In addition, software could also detect error events, such as wrong-side door releases, stopping at the wrong stop mark, and speeding. The same statistics could also be analysed by route, to highlight areas of concern on the infrastructure.

To date, the limited research into the use of OTDR for performance monitoring has been piecemeal and/or unvalidated. A more coherent, scientific approach is necessary to guide future research. The aim of this paper is to present a framework within which to analyse and research OTDR application in human performance monitoring in a more structured fashion than has thus far been achieved. The paper will also discuss the potential use of the signals available in monitoring and improving human performance on the rail network.

OTDR Data

The OTDR record a wide range of signals; only those potentially relevant to driver performance will be discussed in this paper. The data are logged each time one of the monitored signals changes status, i.e. if any signal changes, all signal statuses are logged. The majority of the signals are digital, recording in bitcode format. The exceptions to this are the timestamp on each recording, the distance travelled in kilometres, the latitude and longitude at each recording, the system speed in km/h, and the brake pressure in bar. The distance travelled is an additive signal over the life of the unit (or until reset), so to calculate the distance travelled in terms of a particular journey, the distance value at the start must be known and subtracted from each subsequent recording. The brake and power notches selected by the driver can be derived from the relevant bitcodes. Other driver actions captured include: gear selected (forward, neutral, reverse), headlight selection (none, dipped beam, full beam), use of the horn, door openings (left and right side), and emergency brake activations.

The Irish Rail system, on which the current research has been conducted, features a Continuous Automatic Warning System (CAWS) over parts of the network. The CAWS system provides the driver with an in-cab display of the last signal aspect (colour) encountered. Signal aspects dictate the speeds at which the train can safely travel while still being able to stop at red aspects. The CAWS system also draws the drivers' attention to restrictive aspects, and the driver must acknowledge this auditory warning. The signal aspect according to CAWS is recorded in the OTDR, along with any activations of the acknowledgement button by the driver.

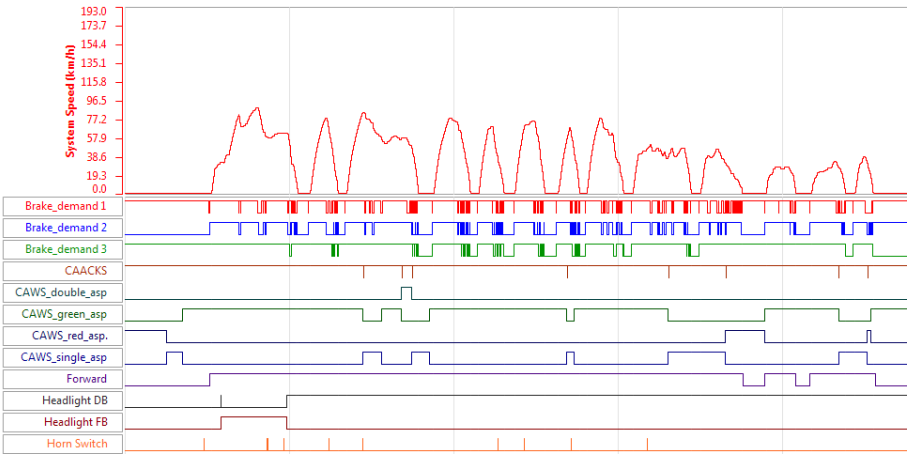


Figure 1. Sample of OTDR Data

Figure 1 illustrates a graphical view of some of the available data showing the train speed, the bitcodes for the brake demand, CAWS aspects and acknowledgements, gear selection and headlight and horn use. The data can be exported to Excel for further analysis.

Analysis Framework

Investigating the potential uses of this data for train and/or infrastructure performance monitoring requires the use of a framework to structure the data analysis. The proposed framework is shown in Figure 2. The overall journey is broken into segments between station stops. Station stops can be reliably identified by those occasions when the recorded train speed is zero and the train doors are opened. The framework then defines six distinct phases within each segment for analysis:

a) Station duties – defined as the time between the doors opening at the station stop (t_1) and the power being applied to leave the station (t_2). Possible metrics of interest in this phase include station dwell times, boarding times, time between doors closing and brake release/power applied, neutral gear selection, and the application of the brakes throughout the stop.
b) Station departures – defined as the time between the power being applied (t_2) and the first speed peak achieved on leaving the station (t_3). Possible metrics of interest in this phase include the power profile (i.e. the specific power notches used when applying power), acceleration rates (several different metrics may be possible), and maximum speed achieved.

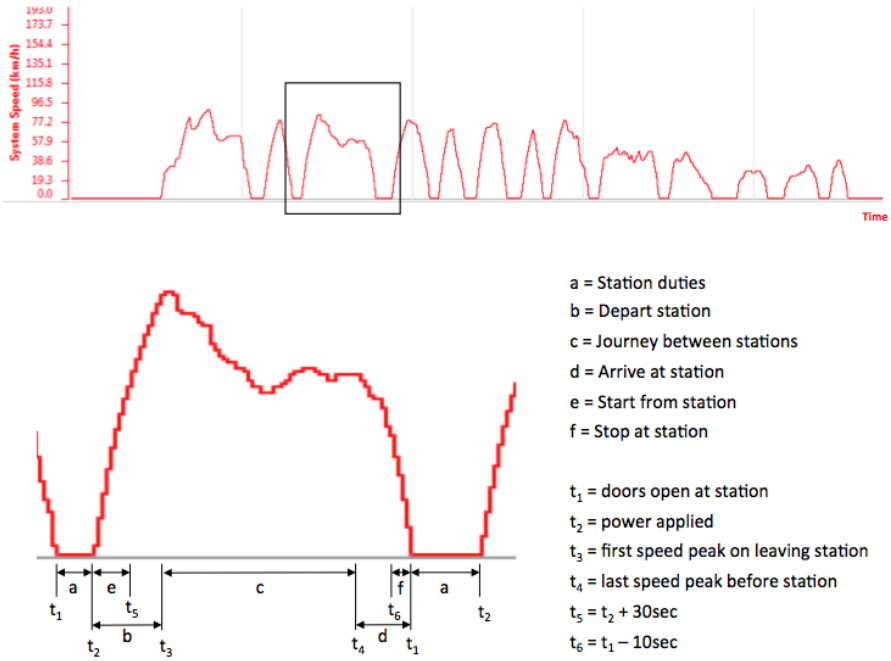


Figure 2. Analysis framework

c) Journey between stations – defined as the time between the first speed peak on departing the station (t_3) and the application of the brakes on approach to the next station (t_4). The application of brakes is analogous to the last peak in the speed profile, and although the application of brakes is not clearly identifiable in the speed profile presented, it is easily identifiable in the raw data. This phase does not occur on short journey segments where braking for the approaching station occurs immediately after the first speed peak but may be quite prolonged on journeys with lengthy intervals between stations. Possible metrics of interest in this phase include speed adherence (where data on line speed is known) and the variation in speed achieved using the median speed and a measure of dispersion.

d) Station arrival – defined as the time from the start of the brake application (t_4) and the doors opening at the station (t_1). Possible metrics of interest in this phase include deceleration rate, braking profiles, maximum brake level applied, and final brake application. More advanced analysis of longitudinal dynamics may also be useful, as well as analysis of power consumption.

The timeframe for station departures and station arrivals may vary widely between different journey segments, due to differences in the distance and permitted speed between the stations. Two further phases are therefore defined to facilitate comparison of departures and arrivals between stations:

e) Station starts – defined as the time between power being applied to leave the station (t_2) and approximately 30 seconds following this (t_5). Metrics of interest are likely to be similar to (b)
f) Station stops – defined as the last ten seconds before (t_6) the doors open at the station (t_1). Metrics of interest are likely to be similar to (d)

Both timeframes (30 seconds and 10 seconds) are arbitrary points, proposed to capture the initial acceleration away from the station and final deceleration towards the station. However, neither timeframe has yet been validated as the most appropriate.

In addition to each of the journey phases, some overall journey statistics may be useful. Speed adherence may be an obvious method to monitor driver performance, but to achieve this with OTDR data would require a model of the network populated with the permitted speeds. This is not currently easily available and speed adherence can be only manually assessed directly from the data. Other metrics from the overall journey include use of the horn, headlight usage, speeds at downgrades to red signals, use of emergency brake, the percentage of time running on restrictive signals, and overall power and brake profiles.

Analysis Potential

The analysis of OTDR data within the defined framework still presents a number of research challenges. First, the data generated is extremely large, with approximately 10,000 lines of data generated for each hour of travel with each signal being logged in each line of data. Big data techniques may be applied, but techniques such as map-reduce (Chu et al., 2007) are unlikely to give the required insights into performance and more complex processing of the data is likely to be required. This requires a computer programming and statistical analysis skill set that is likely beyond most HF practitioners. The data also requires extensive pre-processing to generate information. For example, the braking and power bitcodes must be combined and referenced to generate the actual brake or power selection. Such processing can be relatively easily achieved for a small number of journeys but must be automated in order to facilitate a more comprehensive analysis. The data also suffers from missing signals, such as door closing and wrong recordings, such as errors in the signal aspects generated by the CAWS system and incorrect power levels. These can complicate an automatic analysis and currently, in the research upon which this framework is based, must be manually identified and rectified. Such issues are likely to be temporary, and as the datasets are analysed and the possible human factors applications better identified and understood, more automated data analysis will be possible. But for initial exploration of the data, the processing is necessarily somewhat manual and time-consuming and may require collaboration with computer science and statistics disciplines.

Nevertheless, the data shows strong potential for generating useful knowledge of both train driver and infrastructure performance and should be further researched. Initial research questions include:

What is the variation between drivers and within drivers?

In order to serve as a useful element of a competence management system, the data should be able to differentiate between drivers who are performing well and those who require improvement. The first step in addressing this is to understand the level of variation within and between drivers, to determine whether it is possible to sensitively discriminate between drivers at any level. If it is not possible to discriminate between drivers, or types of drivers, then the use of the data in performance management is likely limited to the monitoring of rule infractions or specific events, e.g. overspeeds, use of emergency brakes, etc. If it is possible to discriminate between driver profiles, additional parameters may be available from the data. Research would therefore be required to identify what parameters can reliably be used to discriminate between good and poor performance.

What are the key parameters that indicate 'good' performance?

In order to facilitate improved performance, it must be possible to differentiate good performance from poor performance. At present there are no validated metrics that differentiate good drivers from poor drivers, or good infrastructure from poor infrastructure. Existing competence management systems may provide a starting point for driver performance to identify initial metrics for measurement and validation where train drivers are currently routinely assessed by supervisors travelling in cab or reviewing a single downloaded train journey to identify deviation from expected performance. More fundamental research is needed to determine metrics for assessing the infrastructure, but models such as the Route Drivability Tool (Hamilton & Clarke, 2005) may provide some insight. The data may facilitate monitoring beyond simple 'red-line' exceedances (e.g. speed adherence) and allow more detailed analysis of driver variability and resilience in line with the Safety II approach (Hollnagel, 2014). Future research should focus on identifying the metrics with the most potential to reveal insights on driver and/or infrastructure performance.

How to incorporate this data in competence management systems?

Once parameters for driver performance have been identified, a further important question is how to incorporate these into competence management systems in an ethical way. The data is collected at a far more granular, detailed and consistent level that is currently possible in driver management systems, and the prospect of continuous monitoring may not be desirable to drivers and may result in increased levels of stress and worry over relatively minor errors or variations in performance. Employee tracking has not been positively received in other organisations (e.g. Amazon) and there are genuine ethical questions about privacy and the use of this data. However, it is already in effect for similar professions (e.g. professional truck drivers). Finding the right balance of safety and performance monitoring is not a trivial task. One possible solution may be to direct much of the data towards the drivers themselves, as in road driving applications (e.g. Albert et al., 2011), to allow them to directly compare their own performance to an average of their peers rather than receiving feedback from supervisors.

Conclusions

The data available from OTDR appears very promising with potential applications in competence management and safety management. The detail available in the data sets may provide an approach to monitoring and improving railway safety performance, based on day-to-day operations. This paper has presented a framework within which to continue research in exploring this potential, and suggested some possible metrics that may be available. Further research is needed to validate these metrics and propose how they can be used in practice to improve railway safety and performance. Within this, there is a need to consider ethical use of the data in supporting drivers and not in 'big brother' style management.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 14/IFB/2717. The author additionally wishes to thank Iarnród Éireann - Irish Rail for their support.

References

Albert, G., Musicant, O., Lotan, T., Toledo, T., & Grimberg, E. (2011). Evaluating changes in the driving behaviour of young drivers a few years after licensure using in-vehicle data recorders. In Proceedings of the *6th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 337-343). Iowa: University of Iowa.

Branton, P. (1979). Investigations into the skills of train-driving. *Ergonomics, 22*, 155-164.

Buksh, A., Sharples, S., Wilson, J.R., Coplestone, A., & Morrisroe, G. (2013). A comparative cognitive task analysis of the different forms of driving in the UK rail system. In N. Dadashi, A. Scott, J.R. Wilson, & A. Mills (Eds.) *Rail Human Factors: Supporting Reliability, Safety, and Cost Reduction* (pp. 173-182). London: Taylor & Francis.

Chu, C., Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G., Ng, A., & Olukotun, K. (2007). Map-reduce for machine learning on multicore. *Advances in neural information processing systems, 19*, 281 – 289.

De Fabris, S., Longo, G., & Medeossi, G. (2008). Automated analysis of train event recorder data to improve micro-simulation models. *WIT Transactions on the Built Environment, 103*, 575 – 583.

Doncaster, N. (2012). "By the seat of their pants" Cues and feedback used by train crew. In J.R. Wilson, A. Mills, T. Clarke, J. Rajan, & N. Dadashi (Eds.) *Rail Human Factors around the World: Impacts on and of People for Successful Rail Operations* (pp. 484-494). Boca Raton, USA: CRC Press.

Donmez, B. Ng Boyle, L., & Lee, J.D. (2008). Mitigating driver distraction with retrospect and current feedback. *Accident Analysis and Prevention, 40*, 776-786.

Dorrian, J., Hussey, F., & Dawson, D. (2007). Train driving efficiency and safety: Examining the cost of fatigue. *Journal of Sleep Research, 16*, 1-11.

Dorrian, J., Roach, G.D., Fletcher, A., & Dawson, D. (2006). The effects of fatigue on train handling during speed restrictions. *Transportation Research Part F, 9*, 243 -257.

Dunn, N., & Williamson, A. (2012). Driving monotonous routes in a train simulator: The effect of task demand on driving performance and subjective experience. *Ergonomics, 55*, 997-1008.

Elliot, A.C., Garner, S.D., & Grimes, E. (2007). The cognitive tasks of the driver: The approach and passage through diverging junctions. In J.R. Wilson, B. Norris, T. Clarke, & A. Mills (Eds.) *People and Rail Systems: Human Factors at the Heart of the Railway* (pp. 115-123). Aldershot: Ashgate.

Evans, A.W. (2011). Fatal train accidents on Europe's railways: 1980–2009. *Accident Analysis & Prevention, 43*, 391–401.

Gillis, I. (2007). Cognitive workload of train drivers. In J.R. Wilson, B. Norris, T. Clarke, & A. Mills (Eds.) People and Rail Systems: Human Factors at the Heart of the Railway (pp. 91-101). Aldershot: Ashgate.

Green, S.R., Barkby, S., Puttock, A., & Craggs, R. (2011). Automatically assessing driver performance using black box OTDR data. In Proceedings of the *5th IET Conference on Railway Condition Monitoring and Non-Destructive Testing* (RCM 2011) (pp. 1-5). New York: Institute of Electrical and Electronics Engineers (IEEE).

Hamilton, W.I. & Clarke, T. (2005). Driver performance modelling and its practical application to railway safety. *Applied Ergonomics, 36*, 661-670.

Hollnagel, E. (2014). Safety-I and Safety-II: The past and future of safety management. Aldershot: Ashgate.

Jenson, M., Wagner, J., & Alexander, K. (2011). Analysis of in-vehicle driver behaviour data for improved safety. *International Journal of Vehicle Safety, 5*, 197 – 212.

Mosimann, C., & Rios, M. (2014). Reduce costs and increase safety by using information generated out of data from black boxes (OTMR, event recorder). In *CORE 2014: Rail Transport for a Vital Economy* (pp. 422-427). Adelaide: Railway Technical Society of Australasia.

Musicant, O., Bar-Gera, H., & Schechtman, E. (2010). Electronic records of undesirable driving events. *Transportation Research Part F, 13*, 71-79.

Naweed, A., & Aitken, J. (2014). Drive a mile in my seat: Signal design from a systems perspective. In Proceedings of *IRSE Australasia Technical Meeting* (pp. 1-7). Victoria: Institution of Railway Signal Engineers Australasia.

Rashidy, E.L., Ahmed, R., & van Gulijk, C. (2016). Driver competence performance indicators using OTMR. In Proceedings of *CIT2016 Congreso de Ingeniería del Transporte* (XII Congress of Transport Engineering) (pp. 354-361). Madrid: Foro de Ingeniería del Transporte.

Strathie, A., & Walker, G. (2015). Can link analysis be applied to identify behavioural patterns in train recorder data? *Human Factors, 58*, 205-217.

Walker, G., & Strathie, A. (2014). Combining human factors methods with transport data recordings. In N. Stanton, S. Landry, G. Di Bucchianic, & A. Vallicelli (Eds.) *Advances in Human Aspects of Transportation: Part 2* (pp. 236 – 245). Louisville: AHFE.

Wilson, J.R., & Norris, B.J. (2005). Rail human factors: Past, present and future. *Applied Ergonomics, 36*, 649-660.

Should the steering wheel rotate? Evaluation of different strategies of steering wheel behaviour regarding controllability and driver acceptance while driving in conditional automated mode

Alexandra König[1], Bernhard Schlag[2], & Julia Drüke[1]
[1]Volkswagen Group, Division Group Research, [2]Technical University of Dresden, Department of Traffic and Transportation Psychology
Germany

So far the steering wheel, as an important interface of Driver-Vehicle-Interaction, is not sufficiently investigated in the field of vehicle automation. At the stage of conditional automated driving it is still unclear, how the steering wheel should behave during the phase of piloted hands-off driving and in take-over situations. A driving simulator study was conducted to evaluate three strategies of steering wheel behaviour (full, 50% reduced, no steering wheel movement) regarding controllability of vehicle guidance and driver acceptance. The scenario based on a requested take-over situation in a curve on a rural road. A total number of 62 subjects (36 female, *M* = 39.2 years, *SD* = 11.1) participated in the study. The full steering wheel behaviour resulted in the slightest deviation from the autopilot steering reference in the take-over situation and was most accepted by the drivers in the take-over situation. In contrast the 50%-reduced steering wheel movement was evaluated as best in the piloted phase. The waiver of steering wheel movement reduced vehicle controllability and driver acceptance. Further research in on-road driving studies is required to investigate the functions of the steering wheel in conditional automated driving in a more realistic context and to validate these results.

Introduction

According to the statements of numerous OEMs and internet companies, like Google, the vision of self-driving vehicles is no longer in unreachable distance (Wachenfeld et al., 2015). Nowadays, Park and Traffic Jam Assistants enable partly automated driving in certain driving situations. The raising automation levels will contribute to a new quality of individual mobility. Conditional automated driving is considered to be particularly beneficial to the driver, as it is based on a guaranteed time reserve for take-overs (Gasser et al., 2013). Since the automated function enables hands-off driving and permanent monitoring is no longer necessary, conditional automated driving reveals new scopes for executing non-driving related

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

activities. In conditional automated driving the driver acts as fall back in case of predictable system limits. That is why the design of the interaction between the driver and the vehicle (human-machine interaction, HMI) is particularly relevant for the safety in conditional automated driving, especially in take-over situations. Such situations are characterized by a shift in the responsibility for the vehicle control between the autopilot and the driver. There is a need for an appropriate HMI in conditional automated driving for resuming control over the vehicle to the driver in various take-over situations. Prior research in the field of conditional automated driving focused on the HMI design of effective take-over requests (Naujoks, Mai & Neukum, 2014), the usability of certain driver-vehicle interfaces for take-over requests (Damböck et al., 2012), the effect of non-driving related tasks on take-over quality (Radlmayr et al., 2014) and the development of design guidelines for a human centred design of take-over situations (Gasser et al., 2013). But there are still open fields of research, for instance the role of the steering wheel in conditional automated driving.

It can be argued that the *steering wheel* represents an important driver-vehicle interface in manual driving giving attention to the fact that the steering wheel serves as operating element for the input of the target course as well as a source of visual and haptic feedback of the actual vehicle trajectory. As conditional automated driving enables hands-off driving, the role of the steering wheel as a salient and intuitive HMI during piloted driving and the take-over situation needs a closer look. It can be argued that keeping full steering wheel movements during conditional automated driving would serve as a source of visual feedback for the driver and would therefore be perceived as transparent and reliable. On the other hand, full steering wheel movements could increase the risk of injuries caused by grasping into the rotating steering wheel. Another disadvantage is the continuous movement that is likely to be experienced as disturbing for non-driving related activities such as reading or texting. Furthermore micro-wheel adjustments may cause a decreased confidence in the automated functions. A complete waiver of steering wheel movements during piloted driving would probably enhance comfort and may decrease the risk of injuries by grasping into the steering wheel. Though the waiver of steering wheel movements would go in hand with the deprivation of visual feedback of the vehicle target heading and would lower their ability to take back control of the vehicle during the take-over situations, especially in curves or turning manoeuvres that result in a dissonance between steering wheel angle and wheel angle.

There is a need for a new definition of the role of the steering wheel in conditional automated driving. The question remains, how the steering wheel should behave during hands-off piloted driving. Should the steering wheel keep rotating or could a total waiver of the steering wheel movement be considered? Google is even considering banning the steering wheel completely out of their fully automated vehicles. To the best of our knowledge no study has given detailed consideration to the functions of the driver-vehicle interface *steering wheel* in the context of conditional automated driving yet.

In the study presented here, an exploratory procedure was chosen to give a first look at the new ground of the steering wheel behaviour in conditional automated driving The focus of the driving simulator study was to evaluate the effects of three steering wheel behaviours on the controllability of vehicle guidance and the driver's

acceptance. The steering wheel behaviour differed in the magnitude of turning motion during the piloted driving: 1) full, 2) 50 % reduced and 3) no steering wheel movement. The three strategies were investigated in 1) conditional automated piloted driving and 2) in a take-over situation in a curve. The aim of the study is to give a first exploratory look on the question, how the steering wheel should behave in conditional automated vehicles to provide a safe and pleasant driving experience.

Method

Strategies of steering wheel behaviour

In the driving simulator study three configurations of steering wheel behaviour were tested which differed in their magnitude of turning motion: 1) full steering wheel movement 2) 50% reduced steering wheel movement and 3) no steering wheel movement (table 1).

Table 1.Comparison of the three steering wheel behaviour strategies regarding configuration, potential advantages and disadvantages

	Full steering wheel movement	50 % reduced steering wheel movement	No steering wheel movement
Configuration	Full steering wheel angles	50% reduced steering wheel angles	No steering wheel movement, steering wheel always in upright position
Potential advantages	• Familiarity • Visual feedback • Trust	• Reduced risk of injuries while grasping the steering wheel • Visual feedback	• Reduced risk of injuries while grasping he steering wheel • Enhanced comfort
Potential disadvantages	• Risk of injuries while grasping the steering wheel • Feeling of insecurity	• Discrepancy of steering wheel and vehicle wheels	• Discrepancy of steering wheel and vehicle wheels • No visual feedback • Passivity of driver

In the strategy with full steering wheel movement the steering wheel kept rotating in piloted driving like in manual driving. As a result the steering wheel angle was always in accordance with the wheel angle.

The 50% reduced steering wheel strategy was indicated by 50 %-reduced steering wheel movement during piloted driving. The steering wheel movement were halved, so that they reflected only a 50% proportion of the real wheel angle movements. As a result the steering translation was more direct in take-over situations, which means that the turning motion of the steering wheel resulted in a larger turning angle of the wheels.

The strategy with no steering wheel movement was characterized by the complete waiver of steering wheel movement during piloted driving. The steering wheel remained in the straight ahead position during all manoeuvres. Therefore, the strategy resulted in a dissonance between the wheel angle and the steering wheel angle in a curve or during a turning manoeuvre.

A pre-test with eight participants from the Division Group Research of Volkswagen AG (woman = 3) showed that a take-over in a curve or during an overtaking manoeuvre was difficult to manage in the condition with no steering wheel movement when the autopilot was instantly deactivated at the moment of first driver contact. To improve the controllability in the take-over situation the autopilot function was faded out over a period of 5 s after the first driver`s contact. This was done to avoid oversteering in the moment of the take-over. The fadeout of the autopilot was applied to each of the steering wheel strategies.

Driving scenario

The three strategies of steering wheel behaviour were tested in two situations on a rural road: 1) hands-off piloted driving and 2) a take-over situation in a curve. The drivers started each ride in manual mode and handed the vehicle control to the autopilot 10 s after they started. The piloted ride took about 4 min (5km). Subsequently, drivers were requested to take over the vehicle control in an S-curve (figure 1). The left curve was poorly visible because of greenery and a hillock at the left side. The autopilot approached the S-curve with a speed of 75 km/h. The road width was 3.65 m. The drivers were informed about the imminent take-over by a countdown while reaching the S-curve. The countdown was presented via a message at a head-up display (“Please take over in 10 s”, (1)). They were encouraged to take over by steering, braking or accelerating just at the moment when the message counted down to zero, as shown in figure 2 (“Please take over”, (2)). By means of the countdown the time reserve for take-over of 10 s was guaranteed, which is recommended by prior research for ensuring a safe take-over in conditional automated driving (Petermann-Stock et al., 2013). Furthermore the countdown was used to ensure that all drivers take over at the same time in the curve. After the take-over situation the autopilot faded out during a time of 5 s (3). The ride was stopped by the study coordinator after the participants completed the curve.

Experimental design and dependent variables

In the study a within-subjects-design was chosen. The order of the three steering wheel strategies was varied, resulting in six permutations via a Latin square. The participants were randomly assigned to one of the six permutations.

For the dimension *driver reaction* the following parameters were used for a plausibility check: 1) mode of take-over, 2) point in time of take-over, 3) occurrence of a braking reaction and 4) time of first braking as shown in table 2. Three modes of take-over were 1) stepping on the gas, 2) braking and 3) steering. The point in time of take-over was defined by the difference between the actual time and the required time of take-over, indicated by the countdown. A take-over at an early point of time, before the countdown counted down to zero, was seen as an indicator for uncertainty of drivers and a decreased controllability of the situation. Point in time of braking was defined by the time when the driver pushed the braking pedal for the first time in the curve.

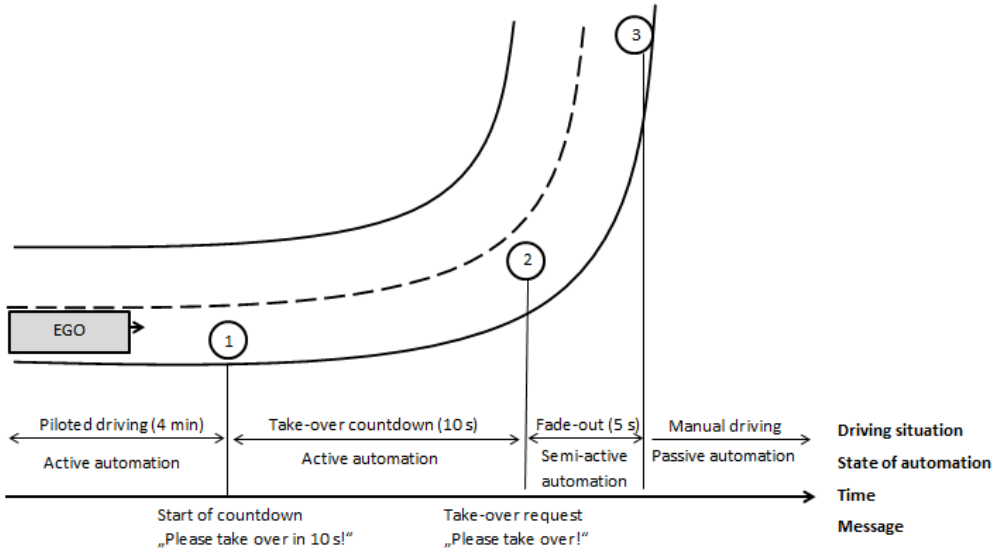


Figure 1. Schematic illustration of the sequence of the take-over situation in the curve.

Figure 2. Screenshot of the driving scenario in the moment of take-over request in the curve ("Please take over!")

Table 2. *Overview of the recorded driving data for plausibility check and measuring controllability of vehicle guidance in the take-over situation*

Dimension	**Parameter**	**Category/unit of measurement**
Driver reaction	Mode of take-over	Accelerating, braking or steering
	Time of take-over	Time before or after requested take-over in seconds
	Braking reaction	Yes/no
	Time of first braking	Time since requested take-over in seconds
Parameter of steering	Root mean square error of steering wheel angle	Degree
Lateral dynamics	Number of exceedances of road markings	Number

Controllability of the vehicle was measured by driving data in the take-over situation and was divided into two parameters: 1) parameters of steering and 2) lateral dynamics. Regarding the parameters of steering the parameter root mean square error of steering wheel angle was analysed (Schmidt, 2009). For this analysis the steering wheel angle was calculated in relation to the autopilot`s steering wheel and was defined as the root mean square deviation from the reference in degrees. For this purpose, a reference ride without take-over was conducted. Under the dimension of lateral dynamics the number of exceedances of the road markings in the curve at the left and the right were recorded. An exceedance was counted as such if distance of the outer back wheel to the road marking exceeded 0 cm (Östlund, Nielssen, Carsten, & Merat, 2004). A negative distance is therefore interpreted as an

exceedance of the road markings. Quantity and duration of exceedances were not considered in the analysis, it was of single interest if the marking was exceeded once to evaluate the controllability of vehicle guidance in the take-over situation. All indicators of controllability were analysed for a time period of 5 s after the required take-over because this was the length of time when the impact of the autopilot faded out.

Driver acceptance of the steering wheel strategies was measured by two questionnaires referring the two situations: 1) the piloted driving and 2) the take-over situation. To compare the effect of the steering wheel strategies between the two situations the items were nearly identical. The questionnaires based on items of the factors trust, comfort, usability and system control of the acceptance questionnaires of driver assistance systems by Arndt (2011) as well as own items. The answer format consisted of 15-point Likert scales based on Heller (1982). Table 3 shows two examples of questions. To gather data about the subjective controllability of the steering wheel strategies the Disturbance Rating Scale by Neukum and Krüger (2003) was used. The scale is clustered in five categories of disturbance rating: 1) not noticeable, 2) noticeable, 3) disturbance of driving, 4) serious disturbance of driving and 5) uncontrollable driving. The categories 2 to 4 are subdivided into three stages, resulting in an 11-point scale.

Furthermore a socio-demographic questionnaire was given to the participants. It included questions about their personal background, annual mileage and experiences with driving simulator studies.

Table 3. Examples of questions of the acceptance questionnaire

How well did you like the steering wheel behaviour during piloted driving?
How well did you like the steering wheel behaviour in the take-over situation?
How comfortable was the steering wheel movement during the piloted driving?
How comfortable was the steering wheel behaviour in the take-over situation?

Participants

A total of 62 drivers (36 female, 26 male) participated in the study. They ranged from 19 to 58 years of age (M = 39.2 years, SD = 11.1 years) and had held their driver's license for on average of 21.0 years (SD = 10.8 years). Drivers were recruited from the participant pool of Volkswagen AG. A majority of the participants already had participated in a driving simulator study and thus were familiar with driving in this simulator. All participants had normal or corrected-to-normal vision.

Experimental setting

For the study, the fixed base driving simulator at Group Research of Volkswagen AG was used (Figure 3). The virtual environment in the simulation was shown on three wide screens (3.00 x 2.25 m), providing a field of vision of about 180°. Furthermore three monitors provided a full rear-view. The driving scenarios were created using the driving simulation software Virtual Test Drive by Vires (Vires, 2014).

Figure 3. Static driving simulator of the Group Research of Volkswagen AG

Procedure

After welcoming the participants they filled in the socio-demographic questionnaire. Then, the participants were introduced to the driving simulator setting. In the beginning, the participants were not informed about the real aim of the study. They were briefed to test a novel function that enables conditional automated driving. During a training session of ten minutes drivers had to activate the autopilot and resume the vehicle control three times after a short piloted ride with full steering wheel movement. This was done to practice the required behaviour regarding the take-over countdown. After the training, the drivers started the first ride with one of the three steering wheel strategies. Afterwards the participants completed the questionnaires concerning the piloted phase and the take-over situation. This procedure was repeated for the other two strategies. At the end, they were briefed about the real aim of the study and were compensated for their participation.

Results

Data analysis

For data analysis the statistical software SPSS 22 was used. Driving data and acceptance rating were analysed by univariate analysis of variance (ANOVA) with repeated measures. Cochran-Q Test was used for analysing the dichotomous dependent variable exceedances of road markings. In case of a violation of the sphericity assumption a Greenhouse-Geisser correction was applied. Post-hoc tests were used to compute pairwise comparisons. The significance level was alpha = 0.05 and was Bonferroni-corrected, if necessary.

Plausibility check

In most of the take-over situations the drivers took over the vehicle by steering(n = 163, 88 %). The remaining take-over manoeuvers were initialized by accelerating (11 %) or braking (1%). The point in time of take-over varied significantly between the steering wheel strategies ($F(2;122) = 11.8$, $p < .001$). While driving without steering wheel movement the drivers took over the vehicle control significantly earlier (M = -0.06 s, SD = 0.11 s, $p_{50\%\text{-}0\%}$ =.001, $p_{100\%\text{-}0\%} < .001$). The result indicates a reduced perceived controllability of the strategy without any steering wheel movement. Most of the drivers applied the brake during the take-over situation in the curve (96.24 %). No significant difference between the steering wheel strategies on the time of first braking were found ($F(2;122) = .331$,$p = .719$).

Controllability

The analysis showed a significant main effect of the steering wheel strategy on the RMSE of the steering wheel angle ($F(1.6;96.2) = 15.6$, $p < .001$). The full steering wheel movement showed a smaller RMSE of steering wheel angle (M = 8.5°,SD = 4.6°) compared to the 50 % reduced steering strategy ($p_{100\%\text{-}50\%} = .009$) and no steering ($p_{100\%\text{-}0\%} < .001$). In the 50 % reduced steering strategy RMSE of steering wheel angle totaled up to 11.1° (SD = 6.6°) and differed significantly from the strategy without steering wheel movements (M = 15.2°, SD = 9.8°; $p_{50\%\text{-}0\%} = .012$).

Cochran-Q Test showed a significant effect of the steering wheel strategy on the number of exceedances of the left road markings in the curve ($X^2(2, 62) = 13.5$,p = .001). Drivers with 50% reduced steering wheel movements exceeded the left road markings less frequent (n = 15), compared to the full steering strategy (n = 28; $p_{100\%\text{-}50\%} = .009$) and the no steering strategy (n = 34; $p_{50\%\text{-}0\%} < .001$, figure 5). The steering wheel strategy showed no significant effect on the number of exceedances of the right road marking ($X^2(2, 62) < .001$, $p = 1.000$). In each steering wheel strategy drivers exceeded the right road markings in 46.8 % of the take-over situations (n = 29).

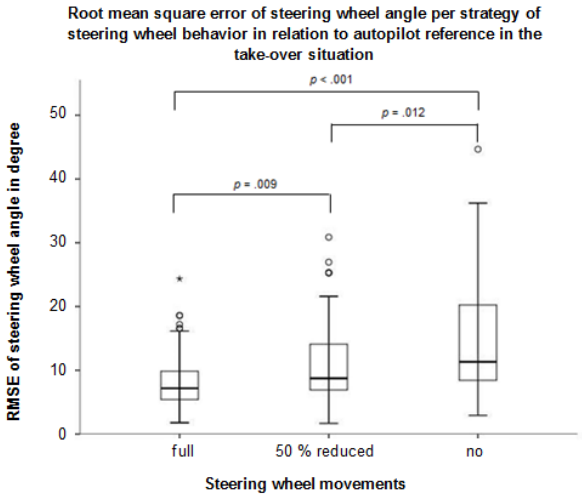


Figure 4. Root mean square error (RMSE) of steering wheel angle per strategies of steering wheel movement shown as boxplot diagrams

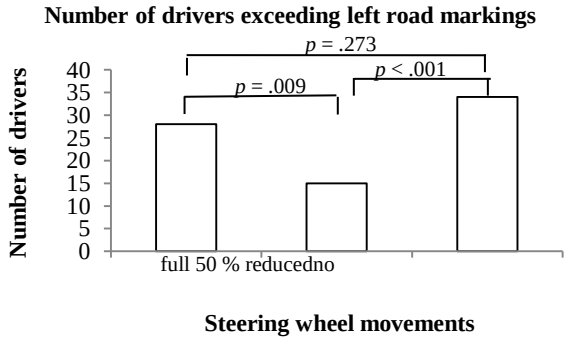


Figure 5. Number of drivers exceeding the left road markings per steering wheel strategy.

Driver Acceptance

Driver acceptance was analysed regarding the two situations of conditional automated driving: 1) piloted driving and 2) take-over situation.

The analysis of the generic evaluation of the three strategies (figure 6) showed a significant effect of the strategies in piloted driving ($F(1.6; 99.5) = 11.3$, $p < .001$) and in the take-over situation ($F(1.8; 110.5) = 30.3$, $p < .001$). In piloted driving full ($M = 10.89$, $SD = 3.1$) and 50% reduced steering wheel movement ($M = 11.73$, $SD = 2.2$) were evaluated better by the drivers than the condition without steering wheel movement ($M = 8.94$, $SD = 4.2$; $p_{100\%\text{-}0\%} = .028$, $p_{50\%\text{-}0\%} < .001$). The comparison of

full and 50% reduced steering wheel movement showed no significant difference in piloted driving ($p_{100\%-50\%}$ = .267). In the take-over situation the drivers evaluated the full steering wheel movement as best (M = 11.60, SD = 2.91) compared to the strategy with 50 % reduced steering wheel movement (M = 10.32, SD = 2.8; $p_{100\%-50\%}$ = .028). The strategy without steering wheel movement were the least accepted by the drivers in the take-over situation (M = 7.27, SD = 3.82;$p_{100\%-0\%}$ < .001, $p_{50\%-0\%}$ < .001).

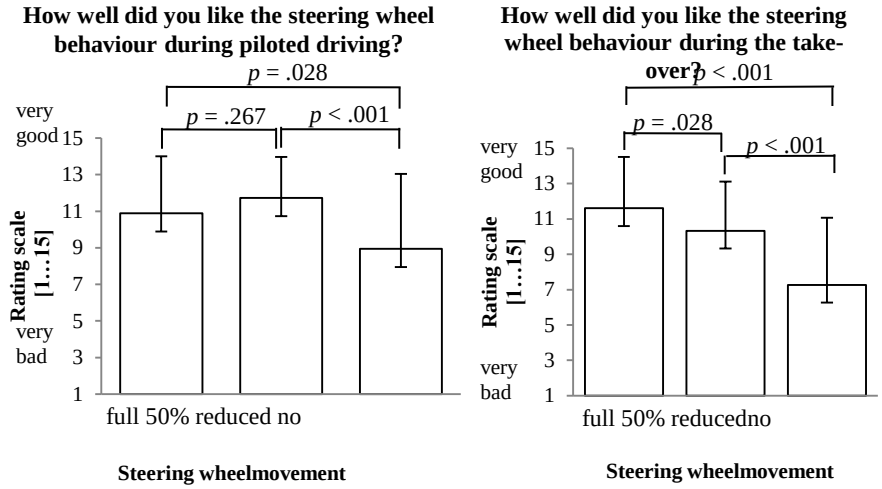


Figure 6. Mean and standard deviation of general evaluation of the three strategies of steering wheel movement during piloted driving (left) and in the take-over situation (right).

With regard to the evaluation of comfort the analysis showed a significant main effect of the steering wheel behaviour on the driver`s assessment of comfort during piloted driving (F(1.6; 98.2) = 9.6, p < .001) and in the take-over situation (F(2; 122) = 24.1, p < .001, figure 7). The strategy with 50% reduced steering wheel movement was rated as the most comfortable by the drivers during the piloted driving (M = 11.9, SD = 2.0, $p_{100\%-50\%}$ = .027, $p_{50\%-0\%}$ < .001) whereas the full steering wheel movement were rated as the most comfortable in the take-over situation (M = 11.6,SD = 2.8, $p_{100\%-50\%}$ = .038; $p_{100\%-0\%}$ < .001). During the piloted phase there was no significant difference in the evaluation of comfort between the full steering wheel movement (M = 10.76, SD = 3.0) and the condition without steering wheel movement (M = 9.42, SD = 4.2; $p_{100\%-0\%}$ = .167). In the take-over situation the strategy without steering wheel movement (M = 7.47, SD = 4.11) was rated significantly less comfortable by the drivers than the other two strategies ($p_{100\%-0\%}$ < .001, $p_{50\%-0\%}$ = .001).

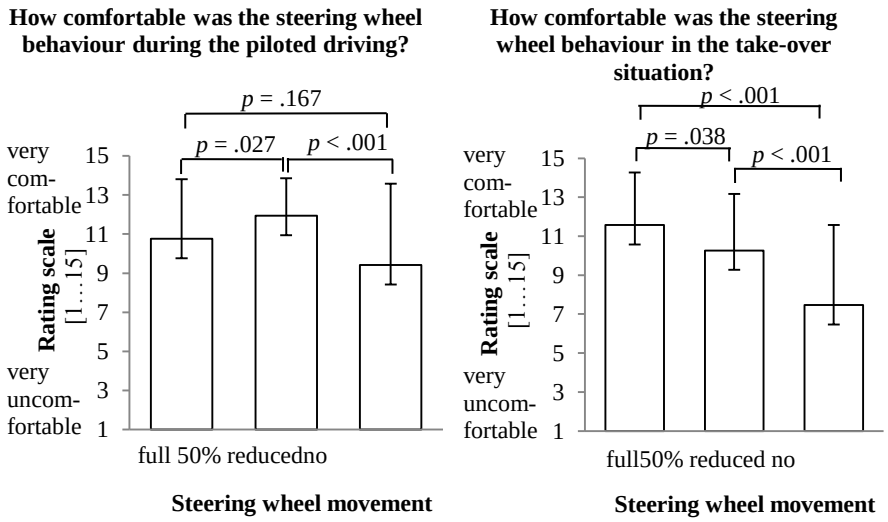


Figure 7.Mean and standard deviation of comfort rating per steering wheel strategy during piloted driving (left) and in take-over situation (right).

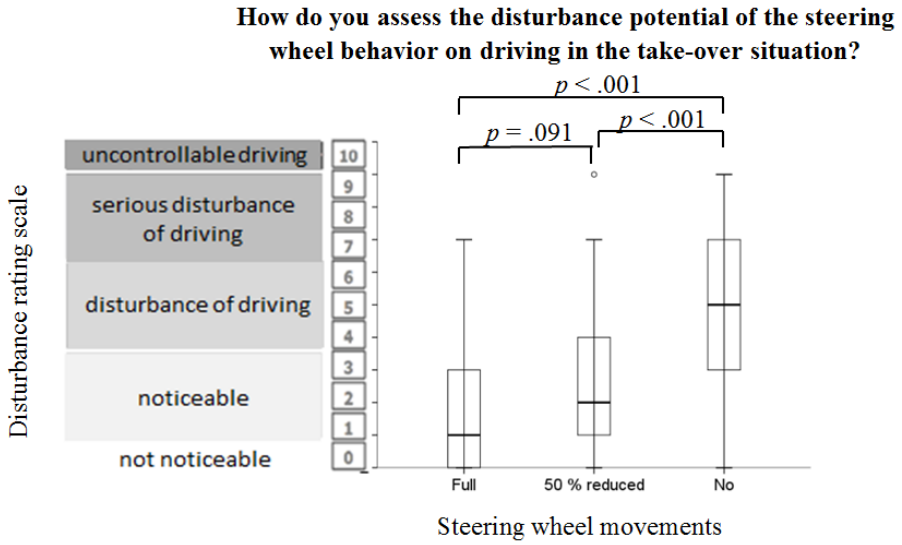


Figure 8. Disturbance rating (Neukum & Krüger, 2003) per steering wheel strategy in the take-over situation shown as boxplot diagrams.

An ANOVA was conducted to evaluate the disturbance rating scale (Neukum & Krüger, 2003) regarding the steering wheel behaviour. Driver`s disturbance ratings differed significantly between the three types of steering wheel behaviour ($F(2; 122)$

= 28.6, $p < .001$, figure 8). The rating of full steering wheel movement ranged from not noticeable to serious disturbance of driving (Range 0-7) whereas 50% reduced and no steering wheel movement ranged up to 9. Nevertheless disturbance ratings did not differ significantly between full and 50 % reduced steering wheel movement ($p_{100\%\text{-}50\%} = .091$). Drivers tend to feel more disturbed when there was no steering wheel movement ($p_{100\%\text{-}0\%} < .001$, $p_{50\%\text{-}0\%} < .001$).

Discussion

The aim of the driving simulator study presented here was to examine the influence of different steering wheel strategies on controllability of vehicle guidance and driver acceptance in conditional automated driving. The results give a first insight into the question, how the steering wheel should behave in hands-off piloted driving and in a safety-critical take-over situation in a curve.

In the take-over situation the full steering wheel strategy resulted in the slightest deviation of steering wheel angles from the autopilot steering reference. This indicates that the continued rotation of the steering wheel enables the driver to take back control over the vehicle in the take-over situations due to a coupling of steering wheel angle and wheel angle. This result is supported by the driver`s subjective rating. The full steering strategy was explicitly preferred by the drivers in the take-over situation.

With the 50 % reduced steering wheel movement the number of drivers exceeding the left road marking in the take-over situation was smaller than with the other two strategies. A possible explanation for this effect could be that drivers overestimated the steering wheel angle of the autopilot in the condition with 50 % reduced steering wheel angles and therefore steered slighter. Especially during piloted driving the 50 % reduced steering wheel movement was evaluated as most comfortable by the drivers.

The no steering strategy caused the highest deviation of the steering wheel angle from the autopilot`s reference in the curve. This result is supported by the lower rating of comfort and higher disturbance potential of the no steering strategy in the take-over situation compared to the other two strategies.

In conclusion, there were strong differences in the effect of the three steering wheel strategies on vehicle controllability and driver acceptance. Overall, the analysis showed a preference of the full steering wheel behaviour in the take-over situation, whereas the 50 % reduced steering wheel movement was preferred during piloted driving. In contrast to the other two strategies, the waiver of steering wheel movement was not well accepted by the drivers, especially in the take-over situation. Thus, a complete waiver of steering wheel movement cannot be recommended. However the results indicate that a reduction of the steering wheel movements could be considered in future conditional automated vehicles, though an adjustment of the take-over process is necessary to improve vehicle control and comfort while resuming control in a take-over situation. This could be done by an enhanced assistance of the autopilot in the take-over situation. Another question that remains unanswered covers the effect of the fadeout of the autopilot in the take-over

situation. Further research is needed to examine the influence of different types and lengths of fadeouts on the controllability of vehicle guidance.

The results of the study should be considered with caution because the study was conducted in a static driving simulator. As Schmidt (2009) and Neukum et al. (2009) showed, lateral acceleration and gear rate of the vehicle are important predictors of the subjective rating of steering wheel functions. Therefore, further research in on-road driving studies is required to examine the functions of the steering wheel in conditional automated driving in a more realistic context to validate the results.

In conclusion, it is not possible to already recommend the use of one of these steering wheel behaviours for future conditional automated vehicles. However, the driving simulator study enables first insight into the role of the steering wheel in conditional automated driving. Paying attention to the restrictions of the study the research question "Should the steering wheel rotate" should be *yes*, but *how* remains as an open question.

References

Damböck, D., Weißgerber, T., Kienle, M., & Bengler, K. (2012). Evaluation of a contact analogue head-up display for conditional automated driving. In N. Stanton (Eds.) *Advances in Human Aspects of Road and Rail Transportation* (pp. 3-12). Boca Raton, USA: CRC Press.

Gasser, T. et al. (2013). *Legal consequences of an increase in vehicle automation*. (Report on the research project F 1100.5409013.01). Bergisch Gladbach, Germany: Bundesanstalt für Straßenwesen.

Naujoks, F., Mai, C., & Neukum, A. (2014). The effect of urgency of take-over requests during conditional automated driving under distraction conditions. In *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE* (pp. 431-438). Krakau, Poland: AHFE

Neukum, A. & Krüger, H.-P. (2003). Fahrerreaktionen bei Lenksystemstörungen – Untersuchungsmethoden und Bewertungskriterien. *VDI -Berichte, 1791,* 297-318

Neukum, A., Leonhard, A., Lübbeke, T., Ufer, E., Krüger, H.-P., Engels, F., & van der Jagd, P. (2009). Fahrer-Fahrzeug-Interaktion bei fehlerhaften Eingriffen eines EPS-Lenksystems. *VDI Berichte . Der Fahrer im 21. Jahrhundert*. 107–124, Düsseldorf, Germany: VDI-Gesellschaft

Petermann-Stock, I., Hackenberg, L., Muhr, T. & Mergl, C. (2013). Wie lange braucht der Fahrer – eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt. In *6. Tagung Fahrerassistenzsysteme. Der Weg zum automatischen Fahren*. München, Germany: TÜV SÜD Akademie GmbH.

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How traffic situations and non-driving related tasks affect the take-over quality in conditional automated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58 (pp. 2063-2067). London, Great Britain: Sage Publications

Schmidt, G. (2009). *Haptische Signale in der Lenkung: Controllability zusätzlicher Lenkmomente*. PhD Thesis, Technical University of Brunswick. Braunschweig, Germany: Deutsches Zentrum für Luft- und Raumfahrt.

Vires (2014). VIRES Virtual Test Drive®. Available from: https://www.vires.com/docs/VIRES_VTD_Details_201403.pdf

Wachenfeld, W., Winner, H., Gerdes, J. C., Lenz, B., Maurer, B., Beiker, S.A., Fraedrich, E. &Winkle, T. (2015). 2. Use-Cases des autonomen Fahrens. In M.Maurer, J. C. Gerdes, B. Lenz, and H. Winner (Eds.). *Autonomes Fahren Technische, rechtliche und gesellschaftliche Aspekte*. (pp. 9-37) Berlin Heidelberg, Germany: Springer.

How does a symmetrical steering wheel transformation influence the take-over process?

Philipp Kerschbaum[1], Kamil Omozik[1], Patrick Wagner[2], Sophie Levin[2], Joachim Hermsdörfer[3], & Klaus Bengler[4]
[1]BMW Group, [2]Former BMW Group Research & Technology,
[3]Department of Sport and Health Sciences, Technical University of Munich,
[4]Department of Mechanical Engineering, Technical University of Munich, Germany

Abstract

During the last several years, numerous studies have been published regarding the human factors challenges in vehicle automation. For conditional automation (SAE, 2014), vehicles must support the driver at non-driving related activities during automated driving as well as manual driving. Both activities have conflicting requirements regarding the vehicle interface, especially for the steering wheel. In order to solve this conflict, Kerschbaum et al. (2015) introduced the idea of changing the steering wheel shape depending on the active driving mode. However, it turned out that the asymmetrical transformation process which is based solely on rotating joints applied in that study prolonged the time drivers needed to take over control from the automation. In this paper, the investigation of an alternative technical solution is presented which allows a symmetric transforming motion of the steering wheel rim. The system was tested in a high fidelity driving simulator study with different take-over scenarios; a state-of-the-art steering wheel served as the control condition. It turned out that the symmetrical transformation indeed influences gaze- and motion reaction parameters, but contrary to the asymmetrical transformation, the main effects are rather positive.

Motivation

In the near future, conditional automation (SAE, 2014) of the driving task is expected to be available for passenger vehicles. At this level of automation, the driver may withdraw from the driving task completely when certain conditions are met. Still, the driver must be ready to take over control should the automation reach a system boundary. In this case, it triggers a take-over request (TOR) while the driver is still provided with a sufficient amount of time to take over. Regarding the human factor of conditional automation, numerous studies have been conducted during the last decade (e.g. see Gold et al., 2013; Lorenz et al., 2014; Merat et al., 2014; Naujoks et al., 2014). Generally, conditional automation requires the vehicle to allow the following three tasks for the driver:

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

- Automated driving while dealing with non-driving related tasks. In case this task involves a display in the vehicle, the driver should ideally have free access to it when seated in physiologically neutral body posture. In this position, comfortable interaction is also possible during long-term automated drives.
- Manual driving. For lateral guidance of the vehicle, the round-shaped steering wheel is commonly applied in modern vehicles.
- The transition between these modes, especially from automated to manual driving if requested by the automation (also referred to as 'take-over situation' in the following). This transition implies the redirection of visual attention, manual regaining of control and taking over based on rebuilt situation awareness (Endsley & Kiris, 1995; Gold et al., 2013)

Automated driving and manual driving result in competing requirements for the vehicle cockpit, especially the steering wheel (cf. Kerschbaum et al., 2015). The static round-shaped rim is a valid option for manual driving. However, it blocks the space right in front of the driver which is valuable for the allocation of displays and controls optimal for non-driving related activities during automated driving. This goal conflict may be solved due to physical transformation of the steering wheel rim when changing the driving mode. This is a salient change of the cockpit geometry, and especially the re-configuration of the circular shape in take-over situations has the potential to accelerate the driver reaction. However, it must not hinder the driver in any way during the transitions between the driving modes.
The experiment reported by Kerschbaum et al. (2015) revealed that an asymmetrical transformation of the steering wheel indeed has influence on the transitions. It decreases the time until drivers start moving after the TOR while it prolongs the take-over time in certain take-over situations. Generally it remained unclear if the results were affected by the unusual asymmetric motion of the rim segments and the investigation of symmetrical transformation seemed relevant in this context. For this reason, a study was conducted in the high fidelity driving simulator of BMW Group Research & Technology focusing on take-over situations. In this study, a symmetrically transforming steering wheel ("TSW") was compared to a state-of-the-art steering wheel (control condition, "CC"). The following research questions need to be answered:

- Is there any influence of a symmetrically re-transforming steering wheel on the guidance of visual attention after a TOR?
- Is there any influence of a symmetrically re-transforming steering wheel on the motoric regaining of control after a TOR?
- Are there any associated effects of a symmetrically re-transforming steering wheel regarding the manual driving task after the transition from automated driving?

Method

Concept and hypotheses

The symmetrically transforming steering wheel was implemented with four rotating joints in the rim, separating it in four segments. The upper segment was equipped with a telescopic element. When driving in automated mode, all upper segments moved

downwards and allowed free prospect to the instrument cluster display. In case of a TOR, they symmetrically moved upwards and form a round-shaped rim as known from modern vehicles within less than one second. During manual driving mode, the steering wheel remained circular and stiff. The process of symmetrical re-transformation during the transition task from automated to manual driving mode is illustrated in Figure 1.

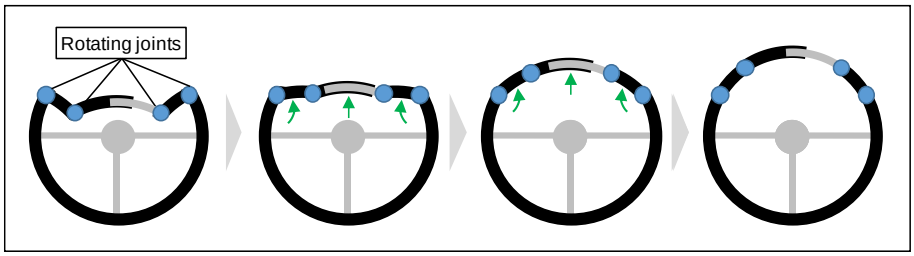


Figure 1. Symmetrical transformation process.

The physical, symmetrical movement of the steering wheel rim segments enhanced the TOR warning signal at take-over situations which was expected to lead to a quicker redirection of visual attention of drivers towards the driving scene. Consequently, the gaze reaction times were believed to be influenced which includes the first saccadic reaction and the time span until drivers have their eyes on the road after the TOR. An additional hypothesis was that the enhanced warning signal influenced the hand movement reaction of drivers. In contrary to the experiment with the asymmetrical transformation, this potentially includes the hands-on times as the symmetrical transformation process is smooth, visually balanced and therefore less distracting during the transition task. It still might shortly attract the driver's attention (Yantis & Jonides, 1984), and make drivers hesitate after their initial start of movement. Regarding the take-over time and the manual driving quality after the transition, no hypothesis could be derived from literature or former experiments. However, the corresponding data was analysed because any detrimental effect would be highly important when discussing the symmetrical transformation of the steering wheel rim. Besides the steering wheel concept, the driver interface of the mock-up vehicle was kept simple: the active driving mode was illustrated by coloured icons in the instrument cluster display. In case of a TOR, a clearly audible tone and a red icon warned the driver to take over control.

Experiment design

In order to ensure the participant's safety, a high fidelity driving simulator was chosen for the experiment which allows participants to experience kinaesthetic feedback while driving. The mock-up car was a 2010 BMW 5 Series equipped with all required mirrors and a full simulated driver interface including visual displays and sound. The original steering wheel of the mock-up car was exchanged with a steering wheel prototype which was able to implement the symmetric transformation of the rim. As illustrated in figure 1, it had four rotating joints integrated and a custom-made telescopic element at the top. This element could change its length along the circular trajectory of the original steering wheel rim. The core structure of

the prototype was made of stainless steel, the next layer was formed by 3d-printed parts. Microcellular rubber was integrated to create a comfortable feeling when touching and grasping the rim. The entire rim was additionally covered with a ductile fabric. The required force for transforming the steering wheel was generated by a linear actuator in the motor compartment and transmitted through the dashboard to the steering wheel by steel cables. Due to this solution, the driver could not hear the actuator noise which was found to be one possible influence on the results reported for the asymmetrical transformation (cf. Kerschbaum et al., 2015). The built-in controller in the actuator was programmed to limit the applied force for safety reasons, the software to control the actuator was implemented on a microcontroller.

The driving simulator allowed participants to drive on a virtual highway with three lanes and one emergency lane. It provided ordinary straight segments and curves. On the virtual highway, event areas were implemented at which specific driving situations could be generated for the driver. Regarding the research questions explained above, two take-over situations were defined. In situation ‘A’, the driver was prompted a TOR while driving on the right lane on straight track (cf. Kerschbaum et al., 2015). At the same moment when the TOR occurred, two crashed cars became visible on the right lane with TTC=7s for the driver to react. The left lane was blocked by traffic cars, so the driver could only swerve to the middle lane or brake after taking over in order to avoid an accident. In situation ‘B’, a TOR was prompted but there was no immanent risk for the driver to collide. The road ahead was free. This situation simulated a TOR due to an internal technical problem of the automation. It was integrated in order to investigate the driver reactions with no extrinsic motivation to react. Additionally, situation ‘T’ (traffic) was implemented. Here, the vehicle approached a traffic jam while driving in a curve. Again, a TOR was triggered with TTC=7s for the driver to brake and avoid a crash. The aim of situation T was to increase the variety of situations and therefore eliminate learning effects. Because no steering maneuver was necessary, situation T was not adduced for data analysis. In order distract participants from the driving scene before the TOR, they were provided with a simple game in all take-over situations. It was displayed on the instrument cluster screen and controlled with a control unit in the middle console of the mockup vehicle. The game popped up at random times, but always before a TOR occurred. The driving situations are depicted in Figure 2.

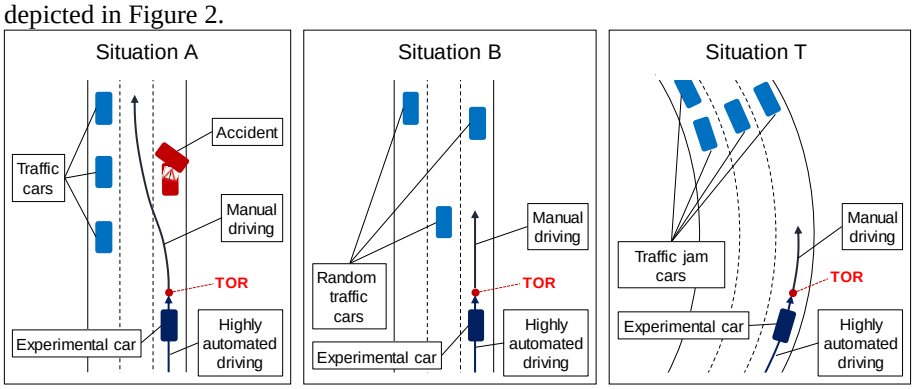


Figure 2. Schemes of the driving situations A (cf. Kerschbaum et al., 2015), B and T.

Regarding the experimental procedure, 62 employees of the BMW Group took part in four consecutive drives. The goal of the first test drive was to familiarize the participants with the driving simulator environment. Only manual driving mode was available with no traffic cars on the highway. In the second test drive, participants could get used to the automated driving mode. With a button on the dashboard, automation could be activated, deactivating was possible by steering, braking or using the button again. On the highway, there were traffic cars and for two times, situation A occurred. This allowed to get participants used to the TOR warning signal and the requirement to engage in manual driving afterwards. In the following, a repeated measures design was implemented with the two within-factors 'steering wheel concept' and 'take-over situation'. Two more experimental drives were executed, one with the transforming steering wheel, one with the control condition concept. In both drives, situation A occurred three times, situation B twice. This distribution was chosen because in situation A, participants were required to engage in a time-critical steering maneuver which was most difficult for the measurement systems to capture without errors. Situation T was not utilized for the comparison of the steering wheel concepts as explained above and therefore interspersed only once. The order of occurrence regarding both factors was permuted to avoid sequence effects. The steering wheel concept which was provided in the first experimental drive was also used in the preceding test drive.

Operationalization

The dependent variables chosen to answer the research questions above are identical to the experiment reported by Kerschbaum et al. (2015). The gaze reaction time (variable t_{gaze}) describes the time span from TOR until the first saccadic reaction of drivers. The road fixation time (variable t_{road}) describes the time span until drivers fixate the driving scene after TOR (see also Gold et al., 2013). Gaze data was recorded with eye-tracking glasses ('Dikablis'). Regarding the driver's hand movement, the first movement reaction (variable t_{move}) and the time span until drivers have their hands on the steering wheel (variable $t_{hands\text{-}on}$) after TOR are calculated. Movement data was recorded using an optical motion capturing system ('VICON' Bonita B10) with seven cameras. Therefore, passive markers were glued onto the drivers' hands and fingers. In order to allow the tracking of the driver's hands even behind the steering wheel, two of the cameras were mounted onto an aluminum frame on the top of the vehicle mockup. They observed the driver's hands through the windshield. The setup is depicted in figure 3.

Figure 3. Motion tracking systems implemented in the mock-up vehicle.

The driving simulator was able to record the driving data of the vehicle, including position on the road, accelerations in lateral (variable a_{lat}) and longitudinal (variable a_{long}) dimension, and manipulation of the brake pedal and steering wheel. Hence, the variables a_{res} and also $t_{take\text{-}over}$ could be derived based on the procedure reported by Gold et al. (2013). In all take-over situations, data recording was started at the moment of the TOR and stopped 20 seconds after the TOR or as soon as the vehicle passed the accident in situation A. For the recurring situations A and B, average values were calculated for the statistical analysis.

Results

Due to technical problems e.g. with either the driving simulator network, driving situations or the vehicle mock-up, 20 driving situations could not be analysed and were excluded. In the following section, the experiment results are reported separately for the guidance of visual attention, motoric regaining of control and the actual take over after the TOR.

Guidance of visual attention

In the take-over situations, the moment of the visual warning signal and warning tone slightly differed from the moment when the steering wheel re-transformed to a circular shape at 28 participants. This issue was due to latency problems within the driving simulator network. The corresponding data sets were excluded from analysis. Figure 4 displays the box-whisker diagrams regarding t_{gaze} and t_{road}.

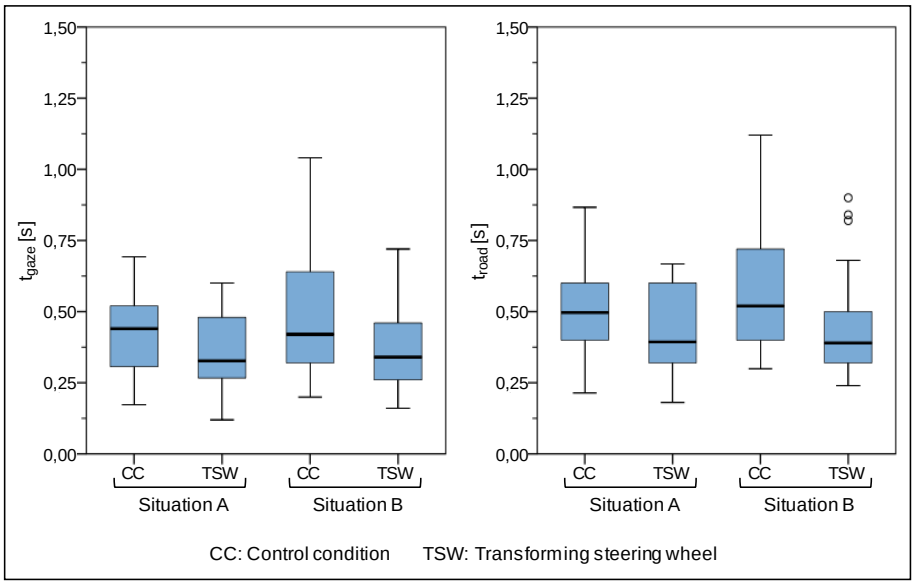


Figure 4. Box-whisker diagrams of t_{gaze} and t_{road}

The diagrams show that the average gaze reaction times are slightly shorter with the symmetrically transforming steering wheel compared to the control condition. This also becomes apparent in the descriptive statistics outline depicted in table 1.

Table 1. Results for t_{gaze} and t_{road}.

Variable	*Situation*	*CC*	*TSW*	*N*
Gaze reaction t_{gaze} [s]	A	M=0.42, SD=0.14	M=0.36, SD=0.14	18
	B	M=0.46, SD=0.21	M=0.38, SD=0.17	
Road fixation t_{road} [s]	A	M=0.51, SD=0.16	M=0.43, SD=0.15	18
	B	M=0.58, SD=0.23	M=0.47, SD=0.20	

A two-way repeated measures analysis of variance was conducted with the data to find out if the differences are statistically significant. The assumption of sphericity was met (Mauchly's test). The analysis revealed a significant effect regarding the steering wheel factor for t_{gaze} ($F(1,17)=8.469$, $p=0.01$, $r=.553$) as well as t_{road} ($F(1,17)=11.848$, $p<0.01$, $r=.641$). Regarding the situation factor, no significant main effects were found for t_{gaze} ($F(1,17)=1.230$, $p=0.283$) and t_{road} ($F(1,17)=2.494$, $p=0.133$).

Motoric regaining of control

During the experiment, several participants lost their passive markers and could not be tracked as planned originally. In several cases, the also turned their hands in a way which did not allow tracking or at least decreased the tracking quality. The corresponding data sets were additionally excluded from analysis, the number of included participants is given in table 2. The box-whisker diagrams of the variables

t_{move} and $t_{hands\text{-}on}$ are depicted in figure 5.

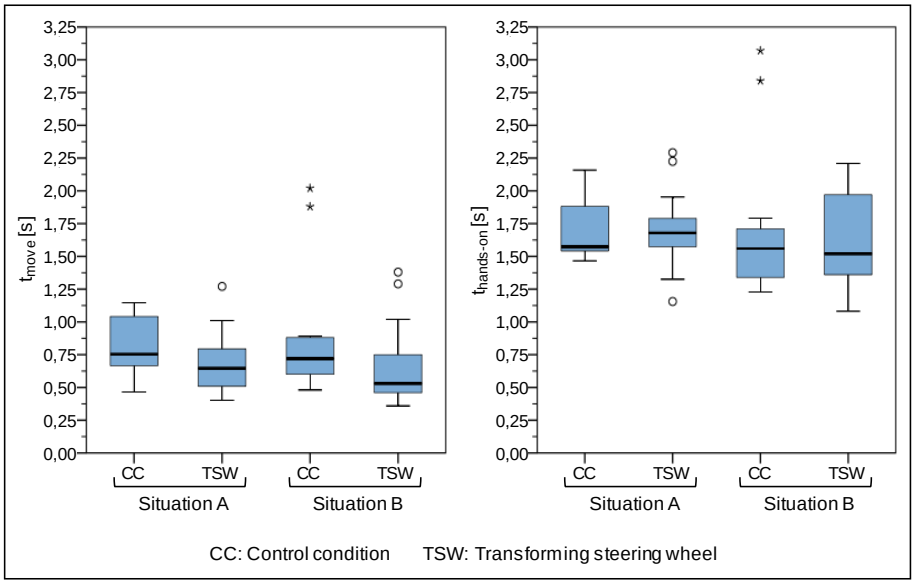


Figure 5. Box-whisker diagrams of t_{move} and $t_{hands\text{-}on}$

Regarding the variable t_{move}, the symmetrically transforming steering wheel led to a slight decrease compared to the control condition as visible in the corresponding box-whisker diagram. Table 2 shows the descriptive statistics.

Table 2. Results for t_{move} and $t_{hands\text{-}on}$.

Variable	*Sit.*	*CC*	*TSW*	*N*
Movement time t_{move} [s]	A	M=0.80, SD=0.22	M=0.70, SD=0.24	14
	B	M=0.87, SD=0.48	M=0.67, SD=0.33	
Hands-on time t_{hands_on} [s]	A	M=1.71, SD=0.24	M=1.70, SD=0.31	14
	B	M=1.70, SD=0.56	M=1.64, SD=0.35	

The assumption of sphericity was met for the data (Mauchly's test). A two-way repeated measures analysis of variance turned out that t_{move} is significantly shorter on average with the transforming steering wheel compared to the control condition ($F(1,13)=4.781$, $p=.048$, $r=.474$). No significant main effect was found for $t_{hands\text{-}on}$ ($F(1,13)=0.185$, $p=.674$). Regarding the situation factor, no significant main effects were found for t_{move} ($F(1,13)=0.106$, $p=.750$) and for $t_{hands\text{-}on}$ ($F(1,13)=0.558$, $p=.468$).

Taking over

Regarding the actual take-over, results of the accelerations and the take-over time are analyzed in the following section. The box-whisker diagram in figure 6 shows that in situation B there are six outliers with high take-over times.

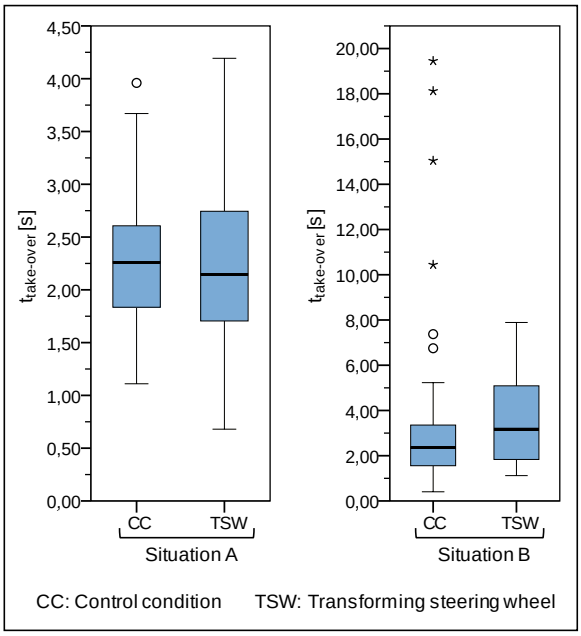


Figure 6. Box-whisker diagram of $t_{take\text{-}over}$.

These outliers are all associated with the control condition concept. The symmetrically transforming steering wheel does not evoke such outliers, all take-over times are below eight seconds. Table 3 gives an overview of mean and standard deviation values.

Table 4. Results for $t_{take\text{-}over}$.

Variable	*Sit.*	*CC*	*TSW*	*N*
Take-over time $t_{take\text{-}over}$ [s]	A	M=2.29, SD=0.66	M=2.24, SD=0.76	48
	B	M=3.56, SD=4.11	M=3.68, SD=1.98	

In a two-way repeated measures analysis of variance, the steering wheel turned out to have no significant effect on the take-over time ($F(1,47)=0.017$, $p=.897$). Because results showed a high difference between the two take-over situations A and B, the influence of the driving situation factor was analyzed in addition. It turned out to be highly significant ($F(1,47)=14.399$, $p<0.001$, $r=.471$). The assumption of sphericity was met for the data (Mauchly's test). The vehicle accelerations during the transition phase after the TOR are illustrated in box-whisker diagrams in figure 7.

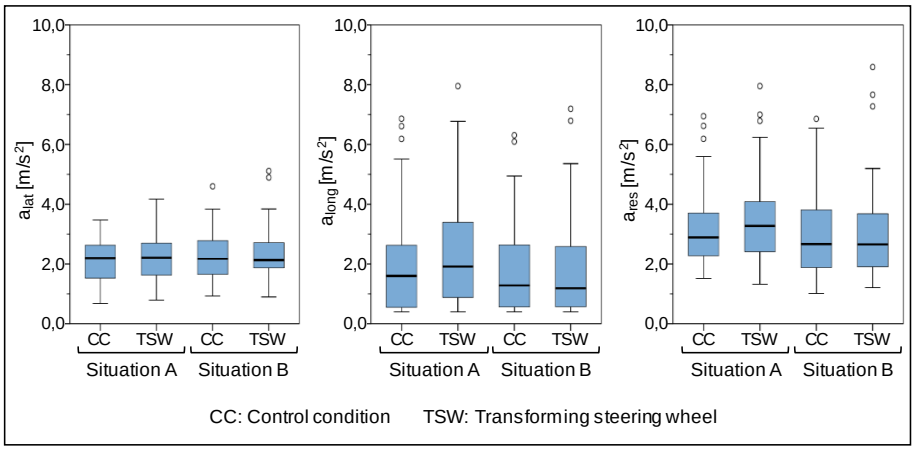


Figure 7. Box-whisker diagrams of a_{lat}, a_{long} and a_{res}.

In the box-whisker diagrams, acceleration levels in lateral and longitudinal dimension are similar. This mainly manifests in the mean and standard deviation values given in table 4.

Table 4. Results for a_{lat}, a_{long} and a_{res}.

Variable	*Sit.*	*CC*	*TSW*	*N*
Max. lateral acceleration a_{lat} [m/s²]	A	M=2.11, SD=0.71	M=2.18, SD=0.74	48
	B	M=2.27, SD=0.81	M=2.28, SD=0.85	
Max. longitudinal acceleration a_{long} [m/s²]	A	M=2.12, SD=1.76	M=2.40, SD=1.86	48
	B	M=1.86, SD=1.63	M=1.87, SD=1.71	
Max. resulting acceleration a_{res} [m/s²]	A	M=3.17, SD=1.31	M=3.42, SD=1.46	48
	B	M=3.02, SD=1.47	M=3.03, SD=1.62	

The two-way repeated measures analysis of variance did not reveal any significant effects. This applies to the lateral acceleration (steering wheel factor: $F(1,47)=0.346$, $p=.559$; situation factor: $F(1,47)=2.791$, $p=.101$), the longitudinal acceleration (steering wheel factor: $F(1,47)=0.614$, $p=.437$, situation factor: $F(1,47)=2.579$, $p=.115$). Hence, neither steering wheel nor situation factor had influence on the resulting acceleration (steering wheel factor: $F(1,47)=0.798$, $p=.376$, situation factor: $F(1,47)=1.548$, $p=.220$).

Discussion

On the contrary to the asymmetrical transformation concept, the symmetrical version has a statistically significant positive effect on the gaze behaviour of drivers in take-over situations. Visual reaction times are shorter regarding both the initial reaction as well as the fixation of the road. Hence, drivers may start to analyse the environment earlier and potentially have more time to react until the vehicle reaches the system limit of the automation.

Similar to the initial visual reaction, drivers also start moving significantly earlier on average with the symmetrically transforming steering wheel. While it was found for the asymmetrical transformation that this time advantage is lost during the transition process and drivers take over later, this is not necessarily the case with the symmetrical transformation. No significant effect could be found here. Hence, the actual realization of transformation indeed changes the way drivers interact with the automation. It acts as an additional, salient cue for participants in take-over situations as it partially accelerates the reaction times. Additional insights specifically regarding the movement reaction might be found due to a detailed analysis of movement velocity profile which has gained acceptance in the corresponding field of research (Hermsdörfer et al., 1996).

Furthermore, there are extreme outliers for the take-over time with the control condition concept in situation B, none with the symmetrically transforming one. In case there is no extrinsic motivation for drivers to take over, some participants seem not to understand the need to take over solely based on the warning signals. The keep interacting with the non-driving related task, unaware of the requirement to take over control. This issue disappears with the transforming steering wheel; it obviously helps drivers to understand the take-over request and to act accordingly.

In conclusion, the results of this experiment indicate that the influence of the steering wheel concept on the take-over process depends on the actual technical concept of transformation. With the symmetrical solution, the main effects are rather positive as they show shorter reaction times shortly after the TOR. Based on these results it may be stated that the transformation principle has the potential to resolve the competing design goals of automated and manual driving, even with positive effects for the transition task. These advantages potentially increase as drivers get used to the transformation concept over time; participants of the reported experiment were completely new to it. Still, it is suggested that the concept should be refined further and simple mechanical solutions need to be found to tap the full potential of steering wheel transformation.

Acknowledgments

The authors thank Alec Ross Fenichel for his support at the experiment, especially for his work regarding the linear actuator and its controllers. Further, they thank Christian Lange for his assistance with the design and realization of the steering wheel prototype.

References

Endsley, M.R., & Kiris, O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37,* 381-394.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop?. In *proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1938-1942). USA: SAGE Publications.

Hermsdörfer, J., Mai, N., Spatt, J., Marquardt, C., Veltkamp, R., & Goldenberg, G. (1996). Kinematic analysis of movement initiation in apraxia. *Brain, 119,* 1575–1586.

Kerschbaum, P., Lorenz, L., & Bengler, K. (2015). A transforming steering wheel for highly automated cars. In *proceedings of the 2015 IEEE Intelligent Vehicles Symposium* (pp. 1287-1292). USA, IEEE Xplore.

Lorenz, L., Kerschbaum, P., & Schumann, J. (2014). Designing take over scenarios for automated driving How does augmented reality support the driver to get back into the loop?. In *proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1681-1685). USA: SAGE Publications.

Merat, N., Jamson, A.H., Lai, F.C., Daly, M., & Carsten, O.M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F*, *27*, 274-282.

Naujoks, F., Mai, C., & Neukum, A. (2014). The effect of urgency of take-over requests during highly automated driving under distraction conditions. In *proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014* (pp. 2099-2106).

Society of Automotive Engineers (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems* (Standard J3016_201401). USA: SAE Society of Automotive Engineers.

Yantis, S. & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 601–621.

Development and evaluation of a method for an intuitive driver's workplace adjustment in a motor vehicle

Yucheng Yang[1], Victor Orlinskiy[2], Ingrid Bubb[1], & Klaus Bengler[1]
[1]Chair of Ergonomics, Technical University of Munich
[2] Bayerische Motoren Werke AG
Germany

Abstract

In a modern vehicle, a driver's workplace permits up to 20 possible adjustments of interior components such as the seat, steering wheel and mirrors. It is difficult and time-consuming for drivers to find the suitable position. However, acceptance of a fully automated adjustment system is low because individual preferences are varied and unpredictable. This has led to the idea of a technical system that leaves drivers in control while at the same time assisting them. This paper develops and evaluates the Intuitive Adjustment System (IAS), through which drivers can adjust the seat, steering wheel and exterior mirrors simultaneously using three parameters on a centralized Human-Machine Interface (HMI). Two of the parameters are for positioning the Hip-point (H-point), the third is to define the individual preference for the torso angle. A method of parameter reduction with an unsupervised machine learning process is proposed, which is applied to a training dataset of individual adjustments and anthropometric records (132 samples). The ascertained H-point patterns generate the ergonomic adjustment strategies and estimate the driver's body height and proportions. The validation experiment (39 participants) shows a positive assessment of the system with better usability and fewer demands on effort compared with the current adjustment system.

Introduction

A typical driver can spend minutes adjusting the driver's workplace and very often only adjusts the basic functions roughly (Sacher, 2009), while a sufficient amount of optimizing potential is ignored (Lorenz, 2011). In addition, due to the large tolerance of humans to body posture, it is very hard to find the ergonomically optimal sitting position and in most cases, poor posture cannot be detected at first sight (Bubb, Bengler, Grünen, & Vollrath, 2015). Zenk (2008) and Lorenz (2011) showed that drivers are not capable of adjusting to an optimal sitting position because of poor instant subjective sensitivity regarding discomfort (Bubb et al., 2015). Therefore, as observed in this study, many drivers find themselves in an iterative, trial-and-error adjustment process.

Three characteristics can explain the problems described above:

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

1. There is a large number of parameters involved in adjusting a driver's workspace (e.g. seat adjustment may have four basic options: longitudinal, height, backrest and inclination)

2. All the switches are to be operated independently. However, their results often have a mutual interaction. For example, the seat longitudinal adjustment (SLA) affects the sitting position, which in turn influences the steering wheel position, and vice versa.

3. The entire adjustment process may include several iterations of switching among multiple adjustment interfaces. For example, the side mirror setting should be corrected after any adjustment of the seat position that affects the eye position. In addition, the HMIs of the current system are located very differently around the driver. Some of them may not be visible in a driving position, e.g. seat adjustments and steering wheel adjustment.

To reduce this complexity, many systems and concepts of seat adjustment in different assistant levels have been developed. The extreme case of simplification is fully automatic adjustment (Mahler, 2006; Durt, Franz, Lein and Zenk, 2009; Breed & DuVall, 1998; Barker & Sakjas, 2010). These systems use various anthropometric parameters that are input manually or by interior sensors. However, the acceptance of fully automatic adjustments is poor because of the large inter-individual and intra-individual variations on sitting postures (Lorenz, 2011). In addition, the input data may be incomplete or corrupted and the acquisition of precise data may lead to more effort for the user or additional system costs.

Partially automatic systems are also being developed by Zenk (2008) and Lorenz (2011), which were better rated in the Lorenz (2011) study. The manual and the automatic aspects of adjustments are mostly divided by functions, which means that some functions, e.g. seat longitudinal adjustment (SLA), are controlled manually, whilst the system optimizes other parameters automatically, e.g. seat inclination adjustment (SIA) based on anthropometric data.

In the above context, this work proposes a new concept for simplifying driver workplace adjustment: the intuitive adjustment system (IAS). It has following features to distinguish it from the existing systems:

1. The control of adjustments of the seat, steering wheel, exterior mirrors is centralized in an HMI with three adjustable parameters.

2. No anthropometric input is required from the user.

3. The driver's body height and proportions are estimated on the basis of the current seat position.

4. No specific sensors are required, except for the position sensors of the adjustable interior components.

5. The adjustment strategy with (in this case) three parameters is learned from the user data. The unsupervised machine learning process may be static or dynamic, depending on the data set.

6. Adjusting any of the extracted parameters influences the estimation of human model, from which other parameters of the driver's workplace can be determined, including the steering wheel and exterior mirrors.

After developing the method, a prototype was built into a BMW 5 series. The new system (IAS) was evaluated against current electric adjustment (TS) with 39 participants. The objective measurements and subjective questionnaires were collected and analysed.

Method

Data set

The development of this concept is based on a learning process with the data which had already been collected for BMW by Technical University of Darmstadt in 2012 (Abendroth, 2012). This experiment demonstrates how customers adjust the driver's workplace to their individual positions without any ergonomic guidance. These individually recorded positions can be influenced by clothing thickness, the type of shoes, injuries to the body, individual preferences, etc.

Human model

The applied human model was derived from the three-dimensional RAMSIS human model and it has been placed in the driver's position by the BMW AG ergonomics department. It is a two-dimension joint model (Figure 1).

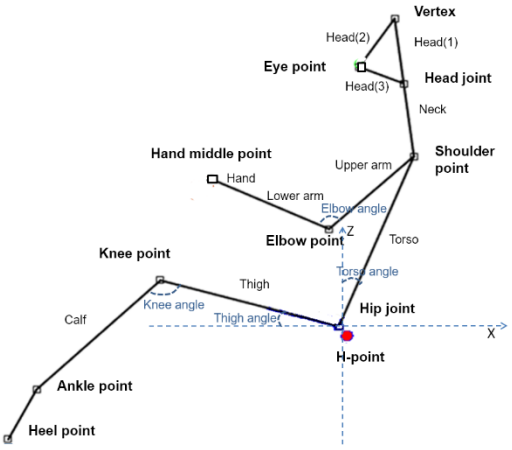


Figure 1. Joints, segments and angles of human model

H-point clustering

Hip-point, also known as H-point, is defined in SAE J1100 (2009) as the pivot centre of the torso and thigh on the two or three-dimensional devices used in defining and measuring vehicle seating accommodation. The H-point can be used as an estimate of the driver's hip joint position from the known seat position and vice versa. In this work, the H-point position is denoted with its x and z-coordinates in

the vehicle coordinates system: Hx, Hz. The H-points of 132 participants are plotted in Figure 2a. The H-point positions, thigh angle and torso angle are calculated from the seat kinematics and position sensor data.

"K-means" as a clustering method has already been identified as a successful learning method (Coates & Ng, 2012). In this work, H-point positions are classified into k mutually exclusive clusters, where k is the number of clusters. Each cluster is composed of its member objects (Hx, Hz) and the centroid (Cx, Cz). K-means iteratively finds a partition in which objects within a cluster are as close to each other as possible, so that the sum of the distances (in this case Euclidean distances) between each centroid and its member points is minimized. This clustering process converges until the distortion function J (Equation 1) is minimized:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \sqrt{\left(H_{x_n} - C_{x_k}\right)^2 + \left(H_{z_n} - C_{z_k}\right)^2} \quad (1)$$

r_{nk} is 1, when the point n is grouped in to cluster k, otherwise it equals 0.

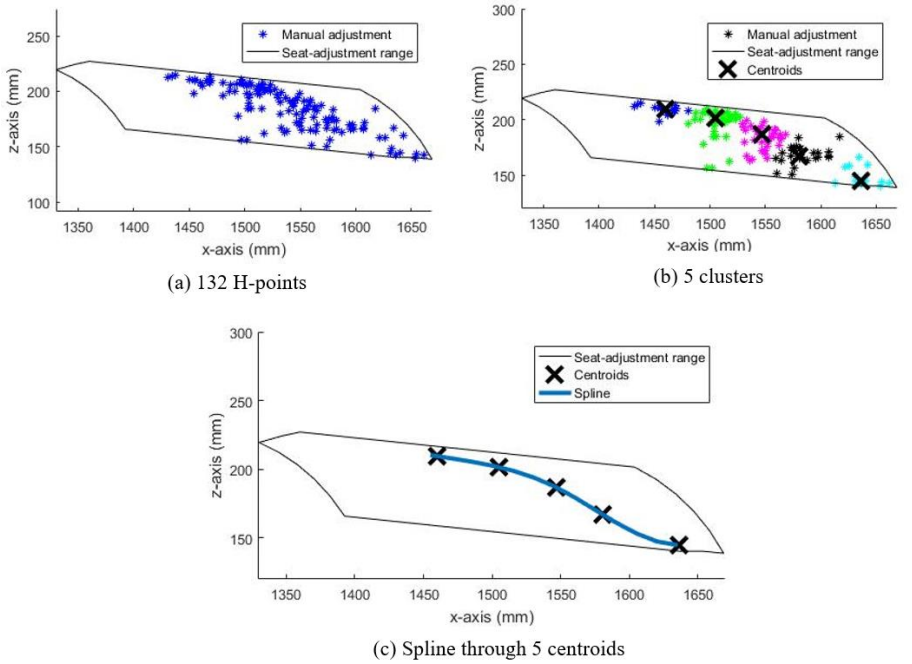
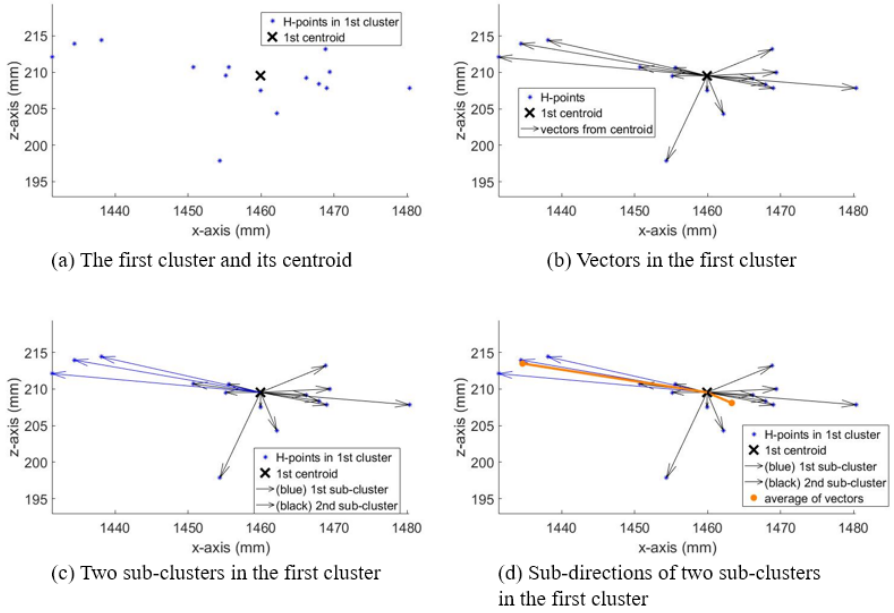


Figure 2. Clustering of H-point

K=5 is chosen in this work taking into consideration anthropometry (5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile), the sample size (132 in total, thus approximately 25 in each cluster), and the optimization of the distortion function [J_{min} = 0 (k=132), J_{max}=7617 (k=1), when k = 5, J=2232, which represents 71% percent of the total reduction of J]. The result is the five clusters shown in Figure 2b. The five clusters are blue, green, magenta, black and cyan. The crosses represent the centroid of each cluster.

Thereafter, cluster centroids are linked together by the cubic spline interpolation in Figure 2c, so that the H-point distribution is simplified to one dimension. The next step is to find out the relationship between individual points and the spline.

H-point sub-clustering



(a) The first cluster and its centroid

(b) Vectors in the first cluster

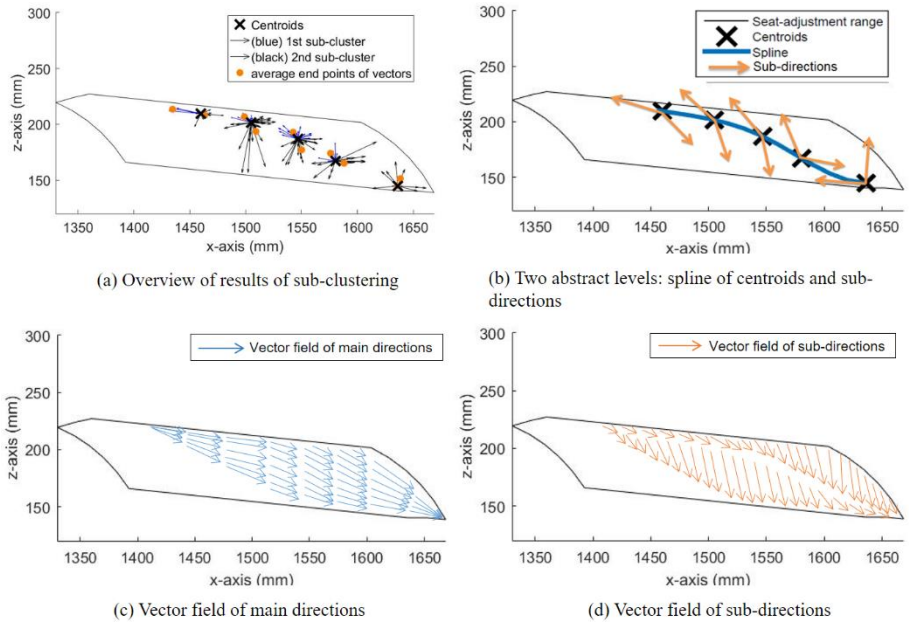
(c) Two sub-clusters in the first cluster

(d) Sub-directions of two sub-clusters in the first cluster

Figure 3. Sub-clustering in 1st cluster

In the first step, the H-point distribution is analysed through k-means clustering and a cubic spline interpolation, providing a one-dimensional description. The goal of the sub-clustering is to analyse the relationship between each centroid and its member points. Since the same four-step sub-clustering method is applied to each of the five clusters, only the process of the first cluster is shown (Figure 3a-d) in detail as an example. Firstly, the 1st cluster is selected, consisting of individual member points and a centroid (Figure 3a). Secondly, vectors are created starting from the centroid and pointing towards each member point. Thirdly, vectors are sub-clustered into two clusters by using k-means (k=2) on their lengths and directions. The 1st sub-cluster is shown in blue and the 2nd sub-cluster in black (Figure 3c). Finally, the average vector within each sub-cluster is defined as the "sub-direction". The average vector means a vector with the average length and the average value of angle to x-axis (+) (Figure 3d).

H-point adjustment pattern



(a) Overview of results of sub-clustering

(b) Two abstract levels: spline of centroids and sub-directions

(c) Vector field of main directions

(d) Vector field of sub-directions

Figure 4. Generalization of main and sub-directions

The same processes of sub-clustering have been applied to the other four clusters so that each cluster has two sub-clusters while one average vector is defined within each sub-cluster as the sub-direction. In Figure 4a, all the five clusters are sub-clustered. The orange points represent the end-points of the average vectors and represent the sub-directions. After having defined two sub-directions in each cluster, individual H-points in the adjustment range are simplified into two abstract levels (Figure 4b): at the cluster level, the spline going through five centroids; at the sub-cluster level, each sub-direction describes the moving direction and the variance from the centroid, which can be interpreted as the relation between the centroid and its member points. Note that the lengths of the sub-directions have been normalized in Figure 4b to provide a better view of the directions. Two vector fields have been constructed to generalize the main direction and sub-direction to other positions in the seat adjustment range: Figure 4c represents the main dimension while Figure 4d represents the sub-dimension.

To reduce the mathematical complexity and implementation effort, adjustments with continuous vector fields are simulated in discrete coordinates. The new adjustment coordinate system (Figure 5) represents the pattern of both vector fields (Figure 4c-d). In the new coordinates, lengths of steps in both directions are equally defined. They are about 1cm in each direction, in order to maintain the consistency of adjustments.

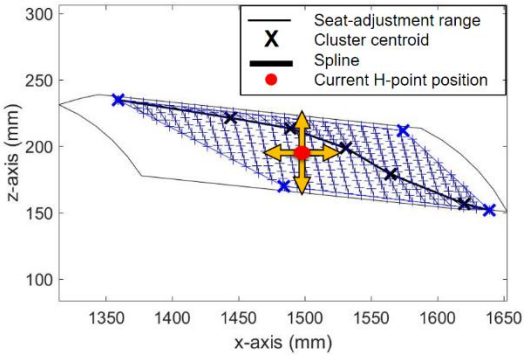
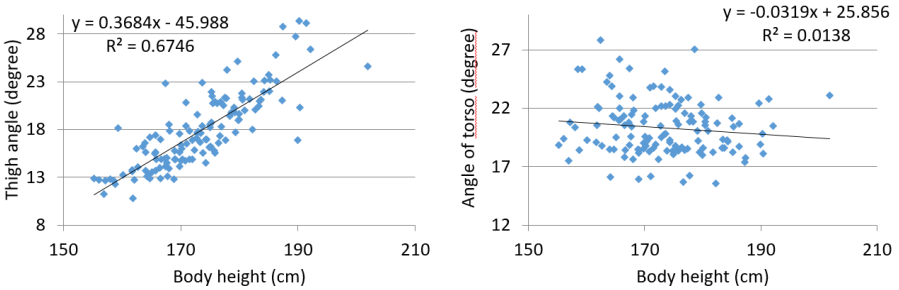


Figure 5. H-point adjustment coordinates

Thigh angles



(a) Linear relation between body height and thigh angle

(b) Angles of torso distribute widely across whole range of height

Figure 6. Linear regression of the thigh/torso angles and body height

The thigh angle is defined as the angle between the thigh and the x-axis (-)[4] (Figure 1). It should be noted here that the thigh angles are not the actual upper-leg angles of the persons sitting on a seat, since body postures can be very different on the same seat surface. The definition of thigh angle assumes that the bottom surface of the thigh is in close contact with the seat cushion. Figure 6a shows that thigh angles increase with corresponding body height (R^2 = 0.6746). The reason could be that taller people have longer legs (thigh and calf) and tend to have a small knee angle in the limited footwell. With the same seat cushion, they need a steeper seat inclination to offer a better support for the thigh. In fact, the four-bar linkage of the seat height mechanism affects the seat tilt angle, since most people do not adjust the seat inclination by themselves. However, the design of the four-bar linkage is deliberate and follows the aforementioned ergonomic aspect.

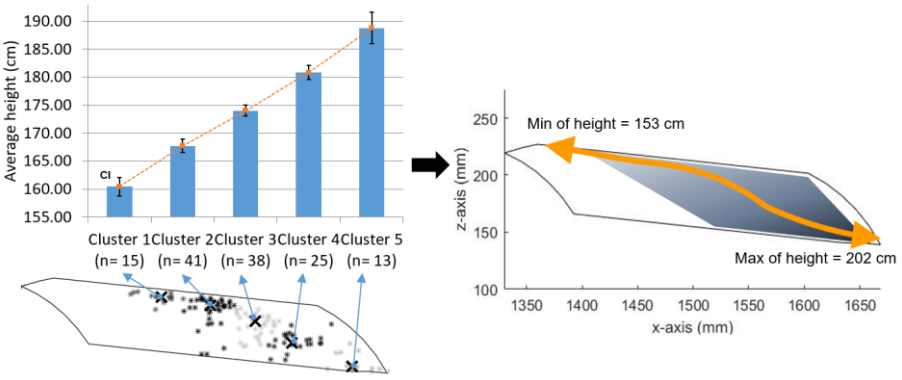
Torso angle

The torso angle is the angle between the torso and z-axis (+) (Figure 1). The definition of torso angle also assumes that the rear surface of the torso is in close

[4] negative x direction

contact with the backrest. Unlike the thigh angles, the correlation between torso angle and body height is not recognizable. The backrest adjustment tends to be very individual. The Figure 6b shows that the torso angles are volatile in every range of body height, though there is a general, vague tendency to decrease. The comfort torso angle according to Bubb (2015) is approx. 25°.

Body height estimation



(a) Increasing body height from 1st to 5th cluster

(b) Body height estimation increases from 153cm to 202cm along or in parallel to the spline

Figure 7. Body height estimation

Body height is an important parameter for the human model. It influences the sizes of different body segments e.g. arm, leg, torso, etc. In Figure 7a, the average body height in each cluster increases consistently from the first to the fifth cluster with margins of error less than 3 cm. T-tests implies that the average height of adjacent clusters differs significantly ($p \leq 0.05$). This indicates that the body height can be estimated, if a given H-point is identified to a cluster. Based on this, a continuous estimation model of body height is suggested. In Figure 7b, the estimated height increases gradually when the H-point travels along the main directions (Figure 4c) towards x-axis plus and z-axis minus. It starts from the minimum 153cm (upper-left corner) to a maximum of 202cm (lower-right corner). Differences in H-point positions can lead to differences in body height estimation, which is an important finding for the human-model-based adjustment method.

Proportion estimation

The proportion is defined as the ratio of seated height to body height (Bubb, 2015), which is a very important factor in the design of the driver's workplace. Under the condition of the same body height, different proportions lead to different leg, torso and arm lengths, as well as eye positions, and so on. Figure 8a shows that the average proportion of individuals in each cluster decreases slightly from the first to the fifth cluster, with a relatively higher fluctuation in comparison with body height. There are only two pairs of clusters that show significant differences in the t-tests: the 1st and 5th, 2nd and 5th, which means that the average proportion decreases from the 1st to the 5th cluster with a certain fluctuation in the middle of the range.

In addition, the manikins from SizeGERMANY show a clear inverse correlation between proportion and body height, under the condition that their proportions are in the same so-called SN (seated normal), SG (seated giants) and SD (seated dwarves) group (Figure 8b).

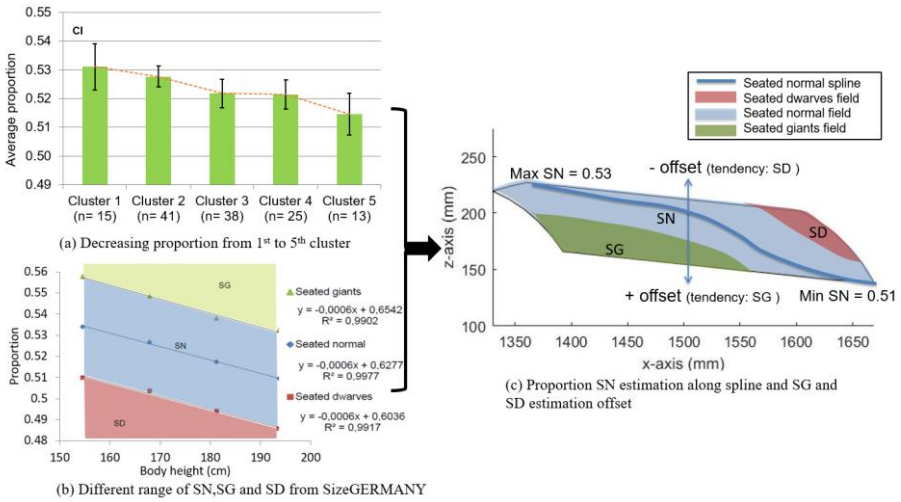


Figure 8. Proportion estimation

Based on the facts above, an estimation model of the 50th percentile proportion (SN) is suggested (Figure 8c): the proportion of the human model will decrease linearly when its H-point travels along the main directions (Figure 4c) towards x-axis plus and z-axis minus, while the estimated body height (Figure 7b) increases. The estimation of proportion starts from the maximum 0.53 to the minimum 0.51. Furthermore, an estimation model of SG and SD is also introduced: as the fine adjustment, an offset on top of SN proportion can change the human model to either SG or SD, depending on whether it is a positive offset or negative offset (Figure 8c). This definition of the offset assumes that every driver intends to ensure a good field of view of the road from their driving position by adjusting the seat height accordingly. Among people with the same body height, those who adjust their seat upwards to ensure a good field of view may have shorter upper bodies and tend to be SDs; on the other hand, those who adjust their seat lower to avoid a collision between the head and the roof may have longer upper bodies and therefore tend to be SGs. The max offset is defined as 0.01, meaning that the proportion estimation of a certain body height is in a range of $p_{SN} \pm 0.01$.

Definition of three parameters

As the result of the data analysis, given an H-point position, the corresponding body height, proportion, thigh angle and torso angle can be derived, which form the 2D human joint model (Figure 1). To control the position of the H-point in a 2D coordinate system, two parameters are necessary. The Parameter 1 (P1) corresponds to the main directions (Figure 4c) of the H-point, which is the result of the clustering and control the H-point to go along, or in parallel to the spline with offset starting

from left-up corner and ending in the bottom right corner in 30 steps evenly. The Parameter 2 (P2) corresponds to the sub-directions (Figure 4d) of H-point, which goes from the upper to the lower boundary of the adjustment field in 19 steps evenly. Torso angles are very individually widely distributed, therefore, the third parameter (P3) is to adjust torso angle. Notice that P1-P3 are not equal to seat adjustment switches, they change the parameter of the human model, which generates all the adjustment for the driver's workspace through reverse kinematics. This system will go through all the adjustments once when the human model is changed and make necessary adjustments for the corresponding posture.

Table 1. definition of adjustment with three parameters

	Value	H-point	Thigh angle	Torso angle	Body height estimation	Proportion estimation
Parameter 1	1-30	Main directions	12°-28°	28°-22°	153cm-202cm	0.53-0.51
Parameter 2	1-19	Sub-directions	-	-	-	Offset to P1: -0.01- +0.01
Parameter 3	1-20	-	-	Offset to P1: -10°-+10°	-	-

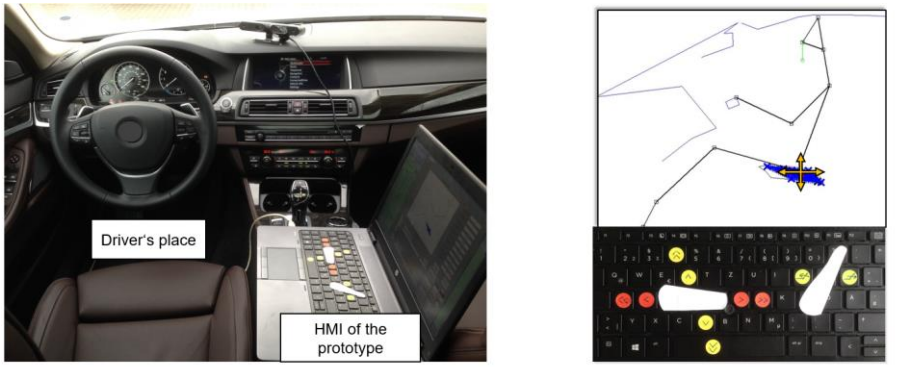
For example, the P3 is more than a backrest adjustment, it changes the torso angle of the human model, which in turn changes the shoulder position, the elbow angle and the eye position. These then affect the adjustments of the steering wheel and exterior mirrors. The functions of three parameters are listed in Table 1.

Implementation of prototype

Figure 9a shows the test environment. The HMI of the prototype (laptop) is placed above the central tunnel, where it can easily be approached and cannot influence the sitting posture. The PC is connected to the On-Board-Diagnose interface (OBD) on the left side of the footwell via an USB/OBD interface cable. The HMI (Figure 9b) and the logic of the prototype is implemented in MATLAB, it communicates with the car through the EDIABAS Tool Set. The red dots on the keyboard represent Parameter 1 while the yellow dots represent Parameter 2, while Parameter 3 is on the right. The "<<" key stands for 2 steps (about 2cm) at once, while "<" is one step (about 1cm). The keyboard is placed in the longitudinal direction of the vehicle so that the pointing direction of Parameter 1 corresponds to the moving direction of the seat. Note that the Ediabas Tool Set, which is used as a tool for diagnosing and fault resolution, can only send one job to the Electronic Control Unit (ECU) at one time through OBD. As a result, after one click by the participants - depending on the corresponding changes of the human model - 12 adjustments (SLA, SHA, SIA, SBA, TSA, HRA, SWL, SWH, MDH, MDV, MCH and MCV[5]) are executed sequentially. The system idle time is approx. 1.5 seconds. Depending on the number

[5] Seat longitudinal Adjustment (SLA); Seat Height Adjustment (SHA); Seat Inclination Adjustment (SIA); Seat Backrest Adjustment (SBA); Thigh Support Adjustment (TSA); Headrest Adjustment (HRA); Steering Wheel Longitude (SWL); Steering Wheel Height (SWH); Mirror Driver's side Horizontal (MDH); Mirror Driver's side Vertical (MDV); Mirror Co-driver's side Horizontal (MCH); Mirror Co-driver's side Vertical (MCV)

of relevant adjustments, the sequence of adjustments takes about 3 to 7 seconds in total.



(a) Test vehicle (b) HMI of the prototype with 3 parameters

Figure 9. the prototype and the HMI

Experiment design

The goal was to compare the new concept with the current electric switches. Independent variables are the intuitive adjustment system (IAS) with 3 parameters and the traditional switches (TS) with 14 parameters (8 parameters for the electric seat, 2 parameters for steering wheel and 4 parameters for mirrors). There are also some additional parameters: body height, proportion and individual preference.

Dependent variables are outcome or response variables and describe what happened as the result of the study (Tullis and Albert, 2013). In this study, there are two kinds of dependent variables: objective measurements and subjective evaluations. There are values for each adjustment parameter, H-point positions, body heights/proportion measurements and estimations, after-task questionaires, NASA TLX (National Aeronautics and Space Administration Task Load Index), and SUS score (System Usability Scale).

First of all, the body height (wearing shoes) and the sitting height of participants are measured in a position where the body is completely upright and stretched. In the familiarization phase, the participants fill out a general questionairre and both IAS and TS systems are introduced. After trying each system, participants use TS (or IAS, the sequence is permuted) to adjust the seat, steering wheel and the mirrors from the initial setting to the individual adjustment respectively and then have a short test drive. Afterwards, they can fine tune or correct their adjustment to the final individual adjustment using the same system. After the operations and positions have been recorded, all the djustments should be again initialized and participants have to get out of the car to remove the motoric memory of the body posture and then get in again to test the IAS (or TS) with the same procedure. Once both systems have been tested, a questionairre and interview session follows.

Results

Descriptive statistics

Thirty-nine employees and students of the BMW Group from different departments, consisting of 13 females (33%) and 26 males (67%) aged between 21 and 60 years took part in and completed the experiment successfully. Twenty-four participants (62%) were between 21 and 30 years old. Thirteen participants (33%) drive more than 20k km/year. Nineteen participants (49%) would not adjust their driving seat during a long-distance trip, though this does not correlate to driving experience. The group (n=11) that drove greater distances (>20k km/year) had significantly clearer individual requirements for each adjustment parameter than the group (n=5) that drove less (<5k km/year).

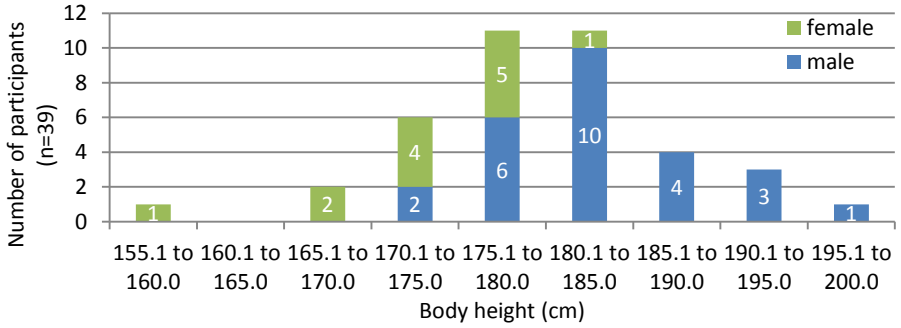


Figure 10. body height distribution of participants

Measurements of body height (wearing shoes) range from a minimum 157.2cm (5th percentile female) to a maximum 196.1cm (95th percentile male), whose distribution (Figure 10) is very close to the normal distribution. Its mean (179.6cm) and median (179.9cm) are also very close.

Subjective evaluation

After-task questions

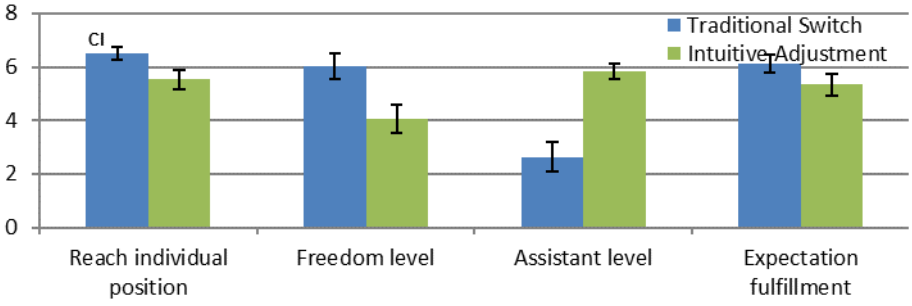


Figure 11. the pre-test questionnaire

Participants answered questions on 7-point Likert scales after each system was tested (Figure 11). T-tests show that there are significant differences ($p \leq 0.01$) between the two systems in each of the four aspects. Most participants think that they reach a better individual position and have more freedom with TS, and that it corresponds better to their expectations. The new IAS offers greater assistance to the paticipants.

NASATLX

Figure 12 shows the comparison of two systems in terms of the different amounts of effort required during use. On average, IAS calls for fewer mental, physical, and temporal demands with a significant level of $p \leq 0.05$ as well as less effort with a significant level of $p \leq 0.1$. In terms of performance, TS may be slightly better with a significance level of $p \leq 0.15$. In addition, they cause almost the same amount of irritation.

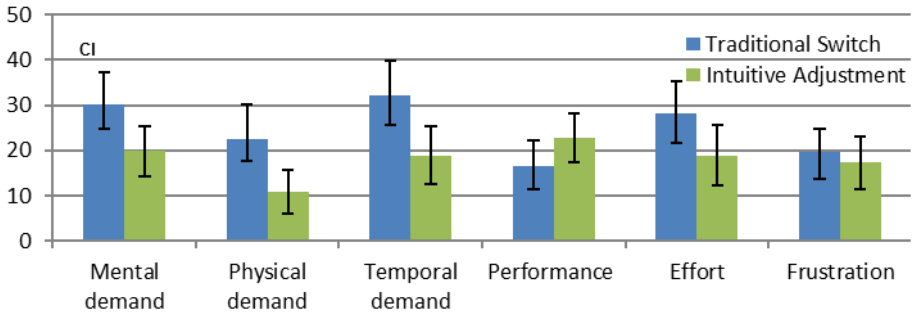


Figure 12. the NASATLX questionnaire

SUS scores

The system IAS has higher average SUS score and a smaller standard deviation than the TS system. The difference between the two scores is very significant ($p \leq 0.01$). The score of the TS system is 71.0, with a confidence interval of ±5.6. According to Bangor et al. (2009), the TS can be classed as "Good". IAS achieves a higher rating of 84.4±4.2, indicating an "Excellent" user interface. An improvement in terms of usability of the new system IAS is clear.

Objective evaluation

Estimation of body height

In the system, body heights of the participants can be estimated when the individual driving positions are reached. Figure 13 shows the estimated body heights over the measured ones with the reference line (y = x). Twenty-three out 39 participants' body height are underestimated, while 16 are overestimated. It should be noted that the measurement of body height is with shoes (i.e. greater than the actual height), which might be a reason of the larger amount of underestimation. $R^2 = 0.6656$ means that the estimations explain about 67% variance of the measurements and they are correlated well. The estimation of body height is linearly controlled by the Parameter 1. Hence, Parameter 1 also correlates well with the body height measurement with the same $R^2 = 0.6656$.

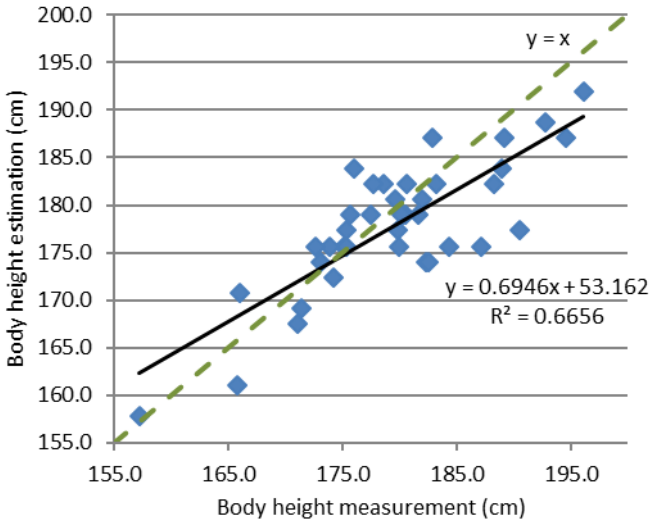


Figure 13. errors of the body height estimation

The study by Willmott & Matsuura (2005) indicates that mean absolute deviation (MAD) is the most natural and unambiguous measurement of average error magnitude. Therefore, the MAD is reported here as the main indication of the estimation quality. In Table 2 the mean of MAD equals 3.82cm, with a standard deviation of 3.08cm, which can be regarded as a small error in relation to body height. A small confidence interval of 0.97 indicates that this result would be similar if it were repeated.

Table 2. statistics of the MAD

	Mean	Standard deviation	Standard error	Margin of error (0.05)
Abs. Δ height (cm)	3.82	3.08	0.49	0.97

Estimation of body proportion

Figure 14 shows the estimated proportion over the measured ones with the reference line (y = x). 28 out 39 participants' proportion are underestimated, while 11 are overestimated. The estimation values correlate badly to the measurements (R^2= 0.078). However, the extreme values of proportion: participant No. 25: SD and No. 9: SG are correctly estimated against SNs. Unfortunately, the number of extreme samples is not enough to prove whether this estimation is stable in the extreme ranges, therefore further studies are needed.

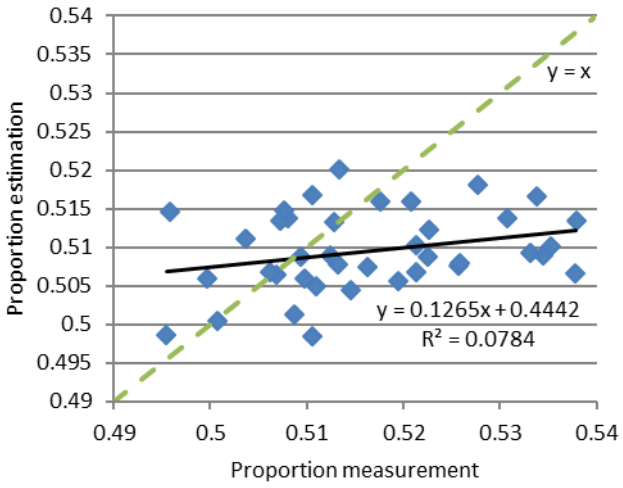


Figure 14. errors of the proportion estimation

Conclusion

In subjective evaluations, IAS is generally very positively rated. It is better than TS in terms of assistant level and usability (SUS test) with less mental, physical and temporal demand. In addition, it requires less effort (NASA TLX). In objective measurements, IAS also reached positions achieved manually by TS in terms of seat, steering wheel and exterior mirror positions with certain acceptable variances, which are comparable to the tolerance of human body (Günzkofer, 2008), while steering wheel longitude adjustments are very individualized. Body height estimation scores well with a small MAD of 3.82cm of the error. The estimation correlated satisfactorily (R^2 = 0.6656) with measurement and Parameter 1 of IAS. Proportion estimation with this sample group of 39 could not be comprehensively rated. Because 37 of 39 are normally proportional as SN, which is a range where individual preference plays the main role in selecting positions. However, two extreme samples of SD and SG were correctly recognized. More reliable conclusions should be made upon future experiments with a larger sample size for SD and SG.

In general, IAS with 3 parameters reaches the intended goal by offering assistance while providing the freedom to make adjustments. Centralized operation prevents the user from switching often and iteratively between several operation interfaces. IAS achieves a positive rating in both subjective and objective evaluation with this PC-based prototype, given the constraints of development time and budget. Therefore, there may be a huge amount of development potential for improvement in terms of implementation and application.

Acknowledgement

The experiment took place at the BMW Group's "Forschungs- und Innovationszentrum", where author Yucheng Yang worked as a paid master thesis

student. Thus, all the data generated is the property of Yucheng Yang and BMW AG. This paper was based on Yucheng Yang's master thesis, which was a part of the master examination and was supervised by Prof. Dr. phil. Bengler and Dipl-Ing. Bubb from the Chair of Ergonomics, Prof. Dr. Pretschner from the Chair of Software Engineering at the Technical University of Munich and Victor Orlinskiy, M.Sc. from BMW AG. The original training data, consisting of 132 data records, were collected by Technical University of Darmstadt for BMW in 2012 (Abendroth B., 2012).

References

Abendroth, B.L.I. (2012). *Analyse der Einstellungsgenauigkeit, Ableitung von Gestaltungsempfehlungen und Ueberpruefung einer weiterentwickelten Version* (unpublished). Technische Universität Darmstadt, IAD.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, *4*, 114–123.

Barker, S.D., & Sakjas, H. (2010). System and method for positioning a vehicle operator. misc, Patents US7830245.

Breed, D.S. and DuVall, W.E. (1998, May 5). Automatic vehicle seat adjuster. Patents US5748473.

Bubb, H. (2015). Menschmodelle. In *Automobilergonomie* (pp. 221–258). Wiesbaden: Springer Fachmedien Wiesbaden.

Bubb, H., Bengler, K., Grünen, R.E., & Vollrath, M. (2015). Automobilergonomie. book, Wiesbaden: Springer Fachmedien Wiesbaden.

Coates, A., & Ng, A. Y. (2012). Learning feature representations with k-means. In Neural Networks: Tricks of the Trade (pp. 561–580). incollection, Springer.

Durt, A., Franz, M., Lein, R., & Zenk, R. (2009, May 14). Verfahren und Vorrichtung zum Einstellen eines Sitzes sowie Sitz. Patent DE102007053119A1.

Günzkofer, F. (2008). Calculation of Driver's Leg Length via Sensor-Guided Seat Positioning (Deplomarbeit). Lehrstuhl für Ergonomie, TUM.

Lorenz, S. (2011). *Assistenzsystem zur Optimierung des Sitzkomforts im Fahrzeug*. (Dissertation). Lehrstuhl für Ergonomie, TUM.

Mahler, W. (2006). Driver seat adjuster for motor vehicle, has pressure sensors to determine person specific parameters e.g. weight, in which actual seat position is adjusted to reference based on parameters and positions of seat components. misc, Patent DE102004062084B3.

Sacher, H. (2009). Gesamtheitliche Analyse des Bedienverhaltens von Fahrzeugfunktionen in der täglichen Nutzung. (Dissertation). Lehrstuhl für Ergonomie, TUM.

SAE International. (2009). J1100: Motor Vehicle Dimensions.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79.

Zenk, R. (2008). Objektivierung des Sitzkomforts und seine automatische Anpassung. (Dissertation). Lehrstuhl für Ergonomie, TUM.

Predicting driver intentions: a study on users' intention to use

Dorothea Langer, André Dettmann, Veit Leonhardt, Timo Pech,
Angelika C. Bullinger, & Gerd Wanielik
Technische Universität Chemnitz
Germany

Abstract

Future driver assistance systems (ADAS) need driver intention prediction. This can help to configure ADAS by (de)activating them in context specific situations where the driver intends to perform a certain action. User's acceptance of such systems is crucial for their usage. In the present study, an algorithm predicted lane change intentions by combining head movement and surrounding information. An automatic turn indicator function made the prediction visible and was used to examine acceptance of such a system. Twenty-one participants drove ten passing manoeuvres, in two manoeuvres activating the indicator manually and in eight manoeuvres with automatically activated indicator. System acceptance was assessed with the Van der Laan-Scale and a questionnaire on Unified Theory of Acceptance and Use of Technology. Additionally, we investigated the activation moment of the indicator as an objective performance measure. Acceptance measures showed intermediate judgements and a low usage intention, each with ample standard deviations. Social influence was the strongest predictor of usage intention whereas performance expectancy and effort expectancy hardly contributed to the explained variance of usage intention. It is concluded that the intention prediction is evaluated mainly sceptically, while also including excited judgements. This result is discussed with regard to the function using it.

Introduction and Review of Literature

Many driver assistance systems are only relevant with certain manoeuvres or in certain situations, e.g. congestion assistance, traffic light assist or blind spot warning. With interconnection and purposeful (de-)activation of different assistance systems, the workload for the driver could be reduced. To allow for this, it is necessary to detect the user intentions as early as possible and to predict his behaviour. This should ensure that information, warnings and especially system interventions do not come into conflict with driver intentions.

According to Michon (1985) the driving task can be subdivided into three hierarchical and interconnected levels in accordance with the respective control processes. The highest level is the Strategic Level which contains the superior

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

planning of a trip based on knowledge, such as route choice or travel time estimation. The level below is the Manoeuvring Level, which comprises rule-based actions regarding interactions with other road users, collision avoidance and adherence to traffic rules. Examples for such actions are turning, lane changing or overtaking. The lowest level is called the Control Level, which is executed highly automated and unconsciously. It involves all basic actions of car handling, such as steering, braking and accelerating. Driver intentions are especially relevant for the two upper levels of the driving task (Kopf, 2005) whereas the control level doesn't require conscious intentions and decisions. Nevertheless, the actions on this level are crucial as indicators for driver behaviour prediction, because driver intent constitutes an unobservable state which only can be estimated on the basis of driver behaviour. Therefore in this paper driver intention detection and behaviour prediction are used synonymously.

Previous approaches of intention detection and behaviour prediction primarily concentrated on the control level and on driver state independent of the driving situation, for example fatigue (e.g. Bekiaris, 2002; Bellet et al., 2009) or attention (e.g. Rauch et al., 2009). But there are well-investigated situations at the manoeuvring level, for example lane changes. Lee et al. (2004) investigated more than 8600 lane changes from a naturalistic driving study. Based on this they propose to classify lane changes into subtypes according to their cause, for example slow lead vehicle, added lane, enter, obstacle or merging vehicle. Additionally for 500 of the lane changes Lee et al. did a detailed analysis of braking, steering, indicating and gaze behaviour as well as data about the vehicles' surroundings. Regarding intention detection these analyses provide important basic insights for the subdivision of lane change types, the relevance of different measurement parameters and the operational sequence of a manoeuvre. Beggiato et al. (2016) also investigated lane changes but specifically in urban traffic. It was shown that gaze patterns play an important role for the detection of certain lane change types. E.g. for lane changes with slow lead vehicle, certain mirror glance patterns were found to be early and robust predictors, but this didn't apply to lane changes due to an added lane. This shows that intention detection on the manoeuvring level should happen situation specific. Nevertheless, Beggiato at al. state that gaze patterns are too ambiguous to serve as the only predictors for intention detection. For example, gaze patterns before some lane changes and turning manoeuvres at crossroads were sometimes similar. Hence for reliable prediction of lane changes, the integration of vehicle parameters as well as data from the vehicles' surroundings is necessary. Therefore the approach of the current study uses basic driver behaviour parameters on Michon's Control Level like gaze patterns to detect driver intentions on the manoeuvring level even before they are put into practice.

But in order to use information about drivers' behaviour for intention detection, a constant observance and recording for example with camera or eye-tracking systems is necessary. Additionally, the intention detection process is invisible to the driver and its result is only perceptible via the automated (de-)activation of a dependent assistance system. In this case the driver doesn't necessarily understand the reason for (de-)activation. This raises the question of the current study if such a system is accepted by drivers. According to Adell (2014), user's acceptance of advanced driver assistance systems is crucial for their usage. She defines acceptance as 'the degree to

which an individual intends to use a system and, when available, to incorporate the system in his/her driving.' (p. 31). According to this definition, acceptance is a behavioural intention to use a system and real usage if the system is available. For investigating the determining factors of this behavioural intention Adell (2014) proposes to use the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003). In this theory, behavioural intention is determined by three factors: performance expectancy, effort expectancy and social influence. Venkatesh et al. define performance expectancy as 'the degree to which an individual believes that using the system will help him/her to attain gains' (p. 447), effort expectancy as 'the degree of ease associated with the use of the system' (p. 450) and social influence as 'the degree to which an individual perceives that important others believe he or she should use the new system' (p. 451). Furthermore, in the model behavioural intention determines usage together with facilitating conditions. Additionally, the UTAUT contains four moderating variables which are sex, age, experience and voluntariness of use.

Only a few expectations regarding system acceptance can be made in advance, because to the moment there is no system on the market to compare with. Effort expectancy is expected to be low, because the intention detection is derived out of natural gaze behaviour. Out of the assumptions of the UTAUT model it can be derived that the performance expectancy and social influence should determine behavioural intention in a positive way, whereas the relation with effort expectancy should be negative. Furthermore it is expected that performance expectancy of the participants correlates with actual performance of the intention detection.

Method

Participants

A total of 23 participants took part in the test drive, but two of them had to be excluded from the analysis due to damaged driving data files. The remaining 21 participants were aged between 25 and 38 years (M = 31.6, SD = 3.5). Nine of them were female. They drove between 1200 and 50000 km per year ((M = 13176 km, SD = 10832, Med = 10000 km). Due to legal and insurance regulations regarding the test car, all participants had to be employees or students of university and to possess a driving licence. They got no compensation.

Materials and procedure

The test vehicle was a VW Touran equipped with radar sensors (front, rear & blind spot), cameras (front, rear, blind spot), Differential GPS and CAN-recording for environment recognition and positioning. Driver' behaviour was recorded with an inside camera and a head tracking API, with which the gaze direction of the driver could be estimated (for details see Pech, Lindner & Wanielik, 2014). The information about number of lanes, and the motion of the test vehicle in relation to other traffic were merged with the driver's gaze patterns to a real time probability estimation of an upcoming lane change (for details see Leonhardt et al., in press). To make this estimation visible to the participant, an automated indicator function was implicated to a Samsung Galaxy S3 Smartphone which was attached to the central instrument

behind the steering wheel. By that the car's indicator was overlaid and replaced by the smartphone indicator application (see Figure 1)

Figure 1. Driver camera and smartphone application for automated indicator function in the test vehicle

A mostly straight part of a public, but quiet street served as test track. The lane markings of the track were partially missing and there were changing amounts of parking cars at the roadside. Despite this drawback, the street was chosen because it was in close proximity to the university campus where the participants came from and it had a dead end which served as starting point. A first car accelerated to a speed of 30 km/h. The participant was driving the test vehicle equipped with the intention detection/automated indicator. When the first vehicle reached a marked point, the participant started too and followed the first car at a speed of 50 km/h. Participants were instructed to overtake if possible, taking account of speed limit and oncoming traffic. After 540 m a parking bay was used to turn and drive back, repeating the follow-and-overtake manoeuvre. Hence the participants drove laps, each consisting of two overtaking manoeuvres at most. Each participant drove 5 laps resulting in a maximum of ten overtaking manoeuvres. If in a lap the traffic situation didn't allow for at least one overtaking manoeuvre, this lap was repeated immediately. Figure 2 shows a sketch of the test procedure including an overview about independent and dependent variables.

Design

Two independent variables were manipulated in a within-subjects design: information about the intention detection system (none/informed) and indicator activation (manual/automatic). At first the participants were not informed about the intention detection. They were told that the purpose of the study was testing the car software. They were instructed to drive the way they always did. After the third lap it was explained to the participants, that the intention detection tries to predict an upcoming lane change by merging gaze patterns and car environment information announcing the detection by indicator activation. During the first two laps participated activated the indicator by hand and the intention prediction was not visible to them (but nevertheless recorded). Dependent variables were the acceptance concepts Usefulness,

Satisfaction and Usage Intention together with indicator activation time as a performance measure. Because of occasional surrounding traffic and parking cars on the roadside, at times participants had to drive unplanned evasive manoeuvres while approaching the standardized overtaking manoeuvre of interest. The tested intention detection algorithm wasn't trained for this kind of situations resulting in random activation of the automated indicator. Due to this and the very short track length, a comprehensive count of false alarms was not possible.

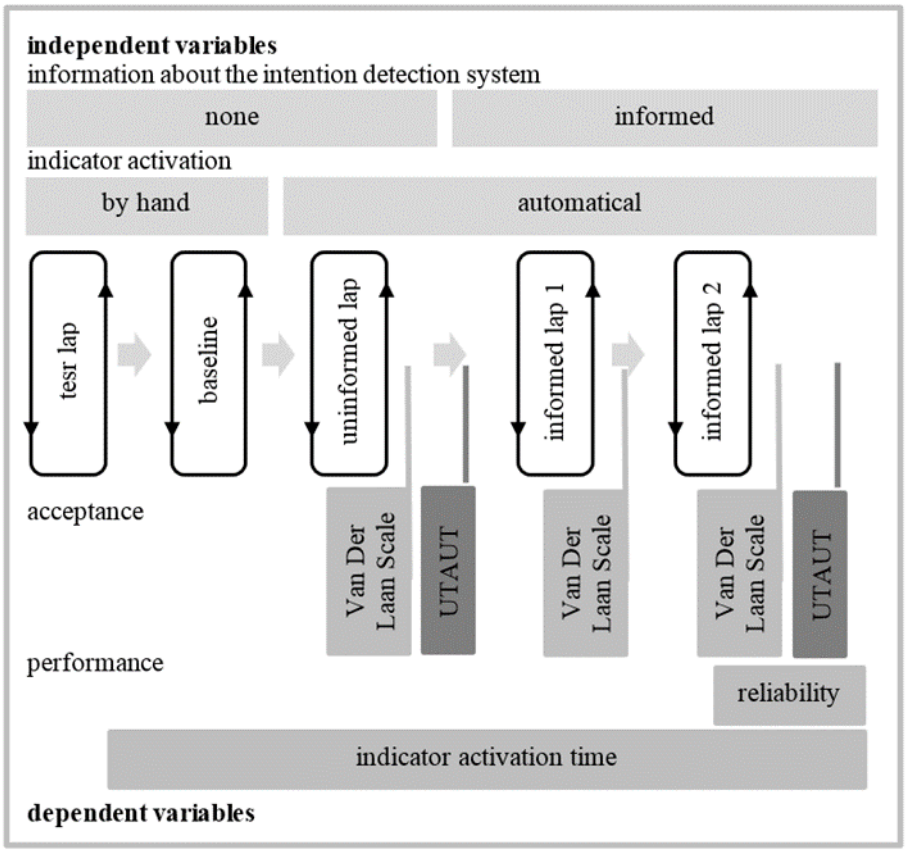


Figure 2. Design and procedure of the test drive.

After every lap with automated indicator signal, the Van-der-Laan scale (Van der Laan, Heino & de Waard, 1997; German version) was used as an economic and general acceptance assessment tool. This scale assesses system acceptance on the two dimensions Usefulness and Satisfying. It contains of nine bipolar ratings in a semantic differential from -2 (e.g. 'bad') to 2 (e.g. 'good'). To get a more detailed understanding of system acceptance, the UTAUT model was assessed additionally after the first and the last lap with automated indicator activation. Because in the current study only a prototype of an intention detection system was available for testing with a restricted user group, the UTAUT concepts usage and facilitating conditions could not be varied and therefore not be investigated. Also of the

moderating factors only gender could be included. Figure 3 shows a schematic picture of the UTAUT model including all determining and moderating relations.

The Items of the UTAUT-Scales behavioural intention to use the system (BI), performance expectancy (PE), effort expectancy (EE) and social influence (SI) in Adells' (2014) adaptation to the driving context were translated to the German language. Each UTAUT Scale consisted of five statements, (e.g. PE1 'I would find the system useful in my driving.') which were rated on a five-point Likert Scale from 'strongly disagree' (1) to 'strongly agree' (5).

The performance of the intention detection was defined as the first indicator activation before an overtaking manoeuvre. Only indicator activation within a maximum time window of eight seconds before lane change was regarded as corresponding to the overtaking manoeuvre because the earliest moment participants activated the indicator was 6.9 seconds before a lane change. Due to missing lane markings on parts of the street, a deviation of the participant's lateral position on the street of one standard deviation from the mean value of all lateral positions of this participant was defined as the lane change moment. This criterion classified trials without overtaking most correctly. Additionally, for seven manoeuvres the lane change moment had to be set using the mean lateral position in the respective lap instead of the mean overall lateral position, because during those manoeuvres parking cars at the roadside narrowed the street considerably.

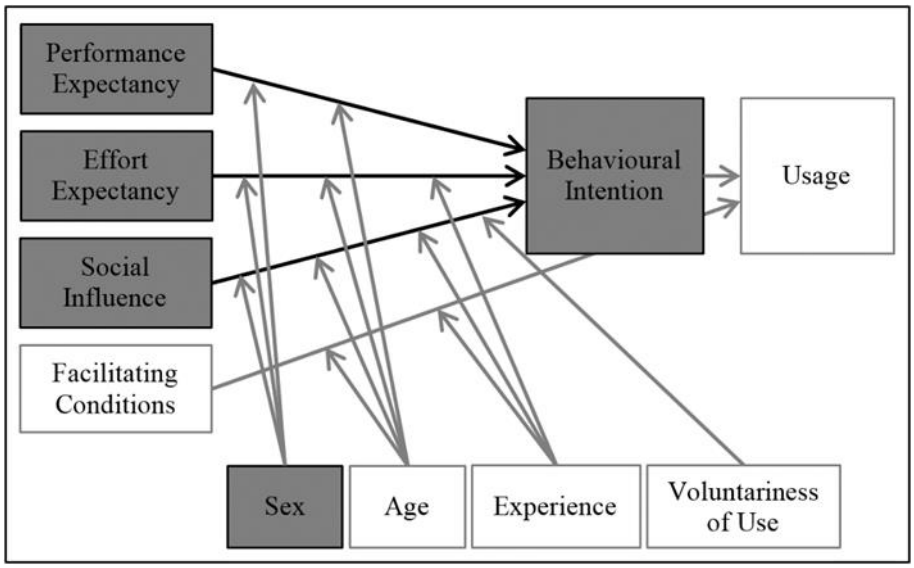


Figure 3. Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003); only the factors marked with grey filling are included in the current study.

Results

Van der Laan Scale

On average, participants showed intermediate judgements regarding usefulness and satisfaction with the intention detection as revealed by automated indicating. But the range of judgements was tremendous, indicating a notable disagreement amongst the participants (see also Figure 4). Table 1 shows the mean scores and standard deviations of the ratings in detail.

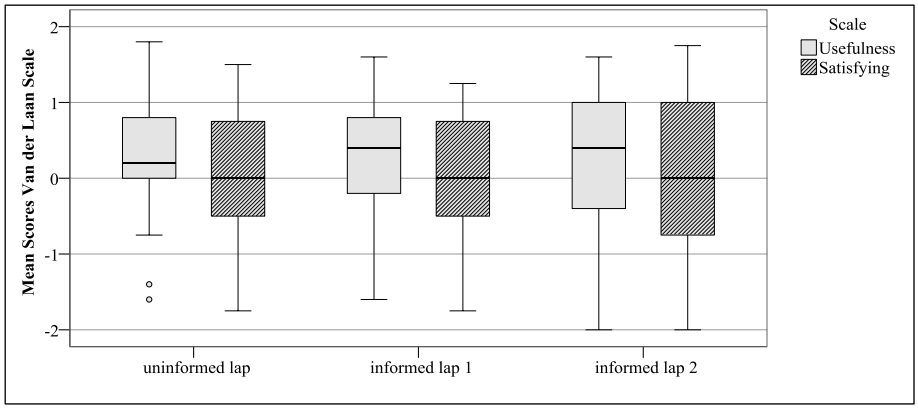


Figure 4. Box-Plot of the acceptance ratings in the Van-der-Laan-Scale over driven laps with automated indicator activation.

UTAUT

The UTAUT model was only assessed after the uninformed lap and the informed lap 2. The mean performance expectancy and social influence ratings were intermediate. The effort expectancy of the participants was rather high. This result contradicts the expectation made before. The core acceptance measure of the UTAUT is Usage Intention, which obtained low mean ratings. But again all standard deviations were large. The detailed values are shown in Table 1.

Table 1. Mean Scores and standard deviations of acceptance ratings after the three laps with automated indicator activation.

	uninformed lap		*informed lap 1*		*informed lap 2*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Usefulness	0.28	0.87	0.26	0.80	0.20	0.98
Satisfying	0.04	0.89	-0.01	0.87	-0.07	1.16
performance expectancy	2.40	0.97	-	-	2.46	1.08
effort expectancy	3.87	0.81	-	-	4.05	0.94
social influence	2.58	0.85	-	-	2.63	0.90
behavioural intention	2.03	1.16	-	-	1.98	1.09

Note: upper two rows: Van-der-Laan Scale (rating -2 to +2), last four rows: UTAUT (rating 1 to 5). N=21.

Behavioural intention is determined by three factors: performance expectancy, effort expectancy and social influence. A regression analysis was calculated to test the assumption of the UTAUT model that behavioural intention is determined by performance expectancy, effort expectancy and social influence. The data of the first assessment violated the data requirements for a regression analysis in several points. Therefore, the regression was only calculated with the data assessed after the informed lap 2. Figure 5 shows some distribution information and the resulting regression scores.

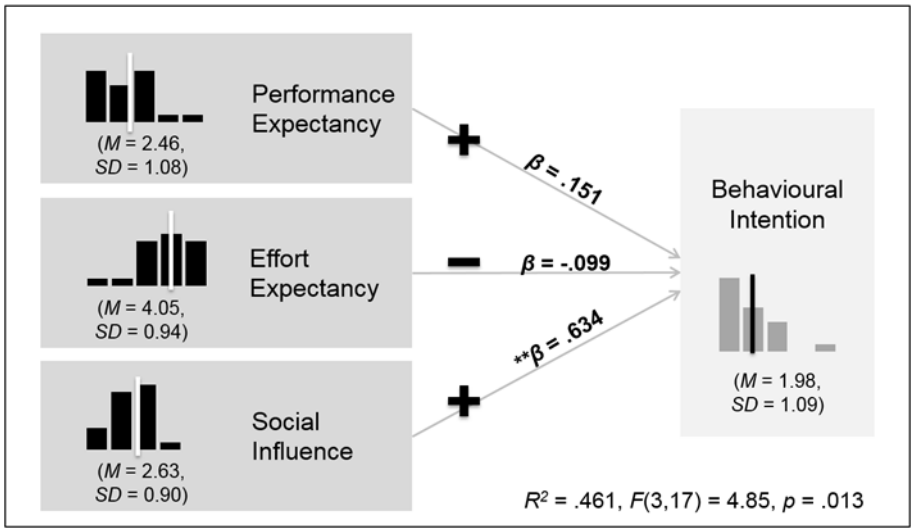


Figure 5. UTAUT model with distribution information of the integrated scales, expected determining relations and resulting regression coefficients.

The beta values of predictors were pointing in the expected direction. Social influence was the strongest predictor of usage intention. Its beta value was more than four times higher than the values of performance expectancy and effort expectancy. The value of their coefficients resulted on comparable levels. The standard errors of the unstandardized regression coefficients of performance expectancy and effort expectancy were larger than those coefficients, indicating an unprecise estimation. Moreover, only the beta value of Social Influence was significant. In total, the regression model explained 46% of the variance (R2 = .461, F(3,17) = 4.85, p = .013). This result was against expectation of the UTAUT model. To account for this result, we tested if the Social Influence mediates the influence of Performance and Effort Expectancy on Behavioural Intention in a mediator analysis (Preacher & Hayes, 2008). Performance Expectancy showed substantial bivariate correlations with Social Norms (r = .48, p = .033) and Behavioural Intention (r = .41, p = .068) whereas the respective correlations of Effort Expectancy were smaller (r = .40, p = .07; r = .22, p = .341). When Social Influence is added as a mediator to the regression, only the correlation of Performance Expectancy with Behavioural Intention decreased significantly (to r = .12, p = .542). The respective 95%-Bias Corrected Confidence Interval (BCCI) ranged from .05 to .85 indicating a significant indirect effect. The correlation of Effort Expectancy with Behaviour Intention

showed no significant decrease (to r = -.07, p = .765; BCCI .17 to .81). The inclusion of gender as moderating variable did not add substantially to the explained variation of the model (R2 = .496, F(7,13) = 1.83, p = .165) and no moderating factor was significant (p between .430 and .837).

Performance

The mean detection time before an overtaking maneuver, when the automated indicator was activated, was 4.85 seconds (SD = 1.22). On average that was more than one second before the mean activation time by hand (Mdiff = 1.32s, SD = 1.08). Additionally, the mean detection time correlated with the rating on the performance expectancy scale (r = -.42, p = .057) even though it didn't become significant. Included into the regression model of the UTAUT measures as supplementary predicting variable, detection time did not add to the explained variance of the model (R2 = .464, F(4,16) = 3.46, p = .032).

Discussion

On average the intention detection, which was made visible with an automated indicator function, is rated as intermediate useful and satisfying. This judgement remains stable, at least over the few usage experiences reported here. But according to Adell's definition, the crucial element of acceptance is usage intention. However, this turns out to be rather low. The UTAUT model suggests three influencing factors to explain behavioural usage intention. Two of them, performance expectancy and effort expectancy, don't show substantial explanatory power to predict behavioural intention. Instead, social influence is the best predictor of the intention to use the intention detection investigated in the current study. This is surprising, especially because actual performance of the intention detection is, as expected, coherent with performance expectancy. Furthermore, effort expectancy turns out to be unexpectedly high, which should diminish the behavioural intention as suggested by the UTAUT. The reason for their weak impact on behavioural intention can only be supposed. One possible explanation could be conceptual blending (Turner & Fauconnier, 2002), a theory on the emergence of new concepts in situations where no established mental patterns are available. An intention detection system is a new and unknown concept so far. There can't be an existing social opinion about the system. Blending theory suggests that in such a case conceptual material from other mental spaces is selected and merged into a new concept or understanding. In this way own precariousness against the system could be blended in, filling the empty space of absent public opinion while Performance and Effort Expectancy can be judged more directly out of the first experience with the system. The mediator analysis supports this assumption at least for Performance Expectancy. But to substantiate this assumption and clarifying its reasons further research is needed.

It can be concluded that the Intention detection is judged mainly sceptical. But the judgements are far from being consistent between people. All results show almost enthusiastic ratings as well as complete rejection of the intention detection system.

This leads to further questions. The first one is if the intention detection is too notional. Participants were asked to rate a highly abstract system which was only

experienced with an artificial indicator function, which isn't useful itself. It is not clear to what extend people are able to differentiate their judgement between the intention detection and the visualization function used, even when asked to do so. Also the potential use of an intention detection system could be insufficiently imaginable to some people. A second, but similar question concerns the extent to which the functioning is understandable to different people. Perhaps the unexpected high ratings of effort expectancy indicate a lack of understanding about the systems functioning. This is underpinned by statements of some participants that they did an extra pronounced head movement for a mirror glance but the indicator wasn't activated. Knowing the system in greater detail it is clear that it uses natural glance patterns over time for prediction. But at least some participants seemed to build simple heuristics on the systems functioning and acting according to them. From diffusion theory it is long known, that especially with complex innovations in the first knowledge phase people require a fundamental understanding of its functioning (Rogers, 1995). Therefore by making the system's principle of operation transparent to participants before testing could help to get more valid results. Beyond that, in further research it should be investigated if imperfect or intransparent automation functions can trigger superstitious behaviour. Also other behavioural adaptations are possible. Therefore long-time experiences with such systems are necessary. Additionally some shortcomings of the current study need to be addressed in further research. With lane change behaviour only a very limited set of situations and manoeuvres were tested here. Furthermore it was only tested in an artificial repeated situation and a small set of highly educated people, which also limits the generalizability of the results.

Nevertheless in the current study for the first time an intention detection system for behavioural prediction on the manoeuvring level was systematically tested for acceptance. With this a first step towards a context specific coordination of advanced driver assistance systems is done. With the obtained knowledge the system can be further developed in a user centred process. Especially when developing an automated function using the intention detection algorithm, e.g. blind spot warning, it seems to be important to thoroughly compile user expectations regarding performance and to make system functioning transparent before testing. In this way future intention detection systems can contribute to more security in driving.

References

Adell, E., Varhelyi, A., & Nilsson, L. (2014). Modelling Acceptance of Driver Assistance Systems: Application of the Unified Theory of Acceptance and Use of Technology. In T. Horberry, A. Stevens, and M.A. Regan (Eds.), *Driver Acceptance of new technology: Theory, Measurement and optimisation* (pp. 23-35). Farnham: Ashgate Publishing, Ltd.

Beggiato, M., Pech, T, Leonhardt, V., Lindner, P., Wanielik, G., Bullinger, A., & Krems, J. (to appear 2017). Lane change prediction: From driver characteristics, maneuver types and glance behavior to a real-time prediction algorithm. In K. Bengler, S. Hoffmann, D. Manstetten, A. Neukum, and J. Drüke (Eds.), *UR:BAN Human Factors in Traffic*. Wiesbaden: Springer Fachmedien GmbH.

Bekiaris, E. (2002) Advanced Driver Monitoring – the AWAKE project. In *e-safety Congress and exhibition proceedings: IT solutions for safety and security in intelligent transport: 16-18 September 2002, Lyon, France*, Vermont South, Australia: ARRB Group Limited.

Bellet, T., Mayenobe, P., Baumann, M., Briest, S., Alonso, M., Vega, M.H., Martín, O., Muhrer, E., Vollrath, M., Minin, L., Montanari, R., Tango, F., & Heers, R. (2009) *Report on the influence of relevant factors (driver characteristics, driver state, driving situation) on driving behaviour and driving errors*. ISI-PADAS Deliverable D.1.1, 117 p. Bron, France: Institut National de Recherche sur les Transports et leur Sécurité (LESCOT: Laboratoire Ergonomie et Sciences Cognitives).

Kopf, M. (2005). Was nützt es dem Fahrer, wenn Fahrerinformations- und -assistenzsysteme etwas über ihn wissen?. In M. Maurer, and C. Stiller (Eds.), *Fahrerassistenzsysteme mit maschineller Wahrnehmung* (pp. 117-140). Berlin: Springer.

Lee, S.E., Wierwille, W.W., & Olden, E.C.B. (2004). A comprehensive examination of naturalistic lane changes. Washington, DC: NHTSA.

Leonhardt, V., Pech, T., Lindner, P., & Wanielik, G. (in press) Fusion of driver behaviour analysis and situation assessment for probabilistic driving manoeuvre prediction. In K. Bengler, S. Hoffmann, D. Manstetten, A. Neukum, and J. Drüke (Eds.), *UR:BAN Human Factors in Traffic*. Wiesbaden: Springer Fachmedien GmbH.

Michon, J.A. (1985). A critical view of driver behavior models: What do we know, what should we do? In: L. Evans and R.C. Schwing (Eds.), *Human Behavior and Traffic Safety* (pp. 485-519). New York: Plenum Press.

Rauch, N., Kaussner, A., Krüger, H.P., Boverie, S., & Flemisch, F. (2009). The importance of driver state assessment within highly automated vehicles. In Conference Proceedings of the 16th ITS World Congress (pp. 1-8). Brussels, Belgium: Intelligent Trasportation Systems and Services for Europe (ERTICO).

Rogers, E.M. (1995). *The Diffusion of Innovations* (pp. 161-203). New York, USA: Free Press.

Pech, T., Lindner, P., & Wanielik, G. (2014). Head-tracking based glance area estimation for driver be-havior modeling during lane change execution. In *17th International Conference on Intelligent Transportation Systems ITSC* (pp. 655 – 660). Red Hook, NY, USA: Curran Associates Inc.

Preacher, K.J. & Hayes, A.F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods 40*, 879-891. doi:10.3758/BRM.40.3.879

Turner, M., & Fauconnier, G. (2002). *The Way We Think. Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research - Part C: Emerging Technologies, 5*, 1-10. German version retrieved from http://www.hfes-europe.org/accept/accept_de.htm.

Driver sleepiness detection based on eye movement evaluation - a driving simulator study

Alina Mashko, Petr Bouchner, & Stanislav Novotný
Czech Technical University in Prague, Faculty of Transportation Sciences
Czech Republic

Abstract

The article provides a study of driver fatigue experimental research on interactive driver simulator. Visually available face features, movements of eyes and facial expression are followed with the help of distant eye tracker. The driver behaviour of sleep deprived participants is observed and compared to that of the rested drivers. This research targets sparing the key features of driver behaviour for further implementation in detection methods of driver fatigue in the human-machine interface of modern and future cars.

Introduction

Automation of many everyday processes is still a challenge – automation of driving is not an exception. Human agent is a soft element in any system. Human body is a system with not only physical properties. The cognitive processes running in human brain while controlling all their physical systems are individual, hard to define and predict. We have been learning to read human body and understand human brain, predict emotions and even imitate humane processes with the help of technology.

Interaction between a driver and a vehicle is a complex system. People have been driving cars for many years. Vehicles, transport infrastructure, cities and regions in general have become smarter, faster and at the same time more disturbed which makes human more vulnerable. Human reaction and response is crucial in the systems and interfaces one is involved. Study of driver behaviour is necessary for their safe interaction with the systems of modern world. Transport systems still depend on input of human agent, especially in-car systems. To improve prediction of driver crucial and dangerous behaviour we study their behaviour in simulated or real environment. Driving simulator experimental studies allow observing human behaviour in critical situations.

Sleep is quite well detected with the help of polysomnography where measurement of brain waves (EEG), eyelid movements (EOG) and muscle tonus (EMG) takes place. The detection of sleep using this technique is quite precise and has significantly contributed to the study of sleep; however because of their invasive nature measurements are only possible in experimental conditions. Non-intrusive measurement of sleepiness is possible with the help of video-based measurements of

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

eye behaviour (blink frequency, duration, time of eye closure/opening, gaze fixations and saccadic movements). Quality of measuring using this method can be compromised by such factors as quality of detectors, lighting situation, or subject wearing glasses. Analysis and evaluation of behavioural sleepiness (body and head movements, face expressions and gestures) is another way of sleepiness detection, where video image analysis or observer rating methods are applied (Anund, 2009).

Our research is aimed at detecting visually available features of driver behaviour in drowsy or fatigued state with the help of tools for visual observation during experiments in a laboratory with driving simulator. The study provides analysis of eye behaviour, simulator obtained data and subjective evaluations. The purpose of research is to combine subjective (self-rating) and objective (eye movements and reaction times) measures for detection of best marginal measure for prevention of falling asleep at wheel.

Experimental measurement of driver fatigue

Research on the problem of driver drowsiness, fatigue and sleepiness is important. Worldwide organizations like National Sleep Foundation in America with project "Drowsy Driving", National Traffic Safety Foundation, or project "Drowsy Driver" in Sweden have been leading extensive research programs in this area.

Experiments with fatigued drivers on driver simulators have been conducted at the faculty the authors are with for tens of years. A complex research has been done to assess driver physiological, psychological states and driving behaviour. Some results are presented in publications by Bouchner et al. (2006).

Experiment platform and measurement method.

Measurement of sleepy driver state is only safe in laboratory conditions. The suggested test is conducted in a laboratory on fixed-base light personal car half-cockpit simulator cut behind the driver seat. The visualization of virtual environment is a 270-degree projection on 3 screens (see Figure 1). Eye movements are recorded with the non-intrusive eye-tracker device Smart Eye Pro.

The driving track used here is a monotonous highway with minimal traffic (Figure 2). There are 9 triggered events (for fresh drivers) and 18 events (for sleepy drivers) in the scenario represented by a special signal that is supposed to engage driver to break till the light changes to green. For this experiment the signal is represented with a lane control red cross/green arrow light placed on roadside gantries (see Figure 3), with green signal set as default and red signal on some of the gantries, where per the given task drivers need to stop. The trigger event serves for measurement a time to reaction and subjective self-evaluation (for sleepy drivers) during experiment. Drivers were instructed to maintain stable speed (around 90 km/h) during the whole experiment.

Participants for the experiment are being recruited mainly among the students and associates from faculty of transportation sciences at Czech Technical University in Prague. Seven people (all male) have participated in the experiment so far (age mean

25.7; SD = 27.6). One of the drivers could not complete the measurement because of simulator sickness, therefore measurements for six participants were available. Due to issues with eye tracker signal for one subject, eye behaviour of 5 drivers was observed. Subjective measure is available for 5 drivers. All participants are active drivers with valid driver's licence, average driving experience 6 years and with around 10 000 km of yearly mileage. Every participant came for measurement twice: once in normal state and a second time after sleep deprivation of at least 24 hours (24 hours since awakening from last sleep). All measurements for sleep deprived drivers took place in the morning hours (with a start around 8-9 am) with most subjects having been awake for the past day and night in a row, while measurements with fresh drivers could happen any time during the day. No energy drinks, excessive coffee intake, strong tea, alcohol or stimulating drugs were allowed during the sleep deprivation period. Simulator drive for fresh participants lasted for about 60 minutes and for sleepy participants – 120 minutes. The 60 minutes' drive is taken as a control state for each participant.

Figure 1. Driver simulator - left and eye-tracking device "Smart Eye pro" - right.

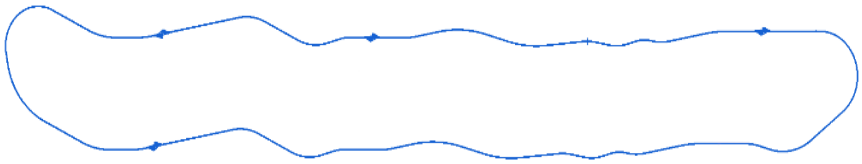


Figure 2. Driving track.

Figure 3. Driving scenario and triggered event realization.

Observation of eye movements

Eye movements may be characterized as saccadic movements being interrupted by fixations, or moving fixations – smooth pursuits (fixation on a moving target). During fixation, the brain is processing the information screened by the eyes. There are studies showing relation between fixation itself and the understanding of screened information during different kinds of activities like reading (Just and Carpenter, 1980), visual search (Hooge & Erkelens, 1997), scene perception (Rayner, 1998) etc. Some of possible measures for fixation are duration time, direction, or position, velocity. Evaluation of driver behaviour by means of measuring gaze duration is described in standard ISO 15007. Eye behaviour presented in this chapter relates to one experiment participant. Behaviour of drivers involved in the experiment was quite individual, besides, the number of participants of this experiment is small (at this stage) and therefore an expert approach was chosen for assessment and analysis of measured data.

Normally, fixation duration may last from 100 ms to over a second (Bergstrand, 2008). Eye behaviour has been followed in our sleepiness research, fixations and blinks analyses has been done over one hour for each 5 minutes' time section starting from the start of experiment. By fixation here a gaze intersection with front screen simulator projection (front scenery view) and with instrument cluster is understood. Analysis of one of the study participants is provided here. It has been observed that number of fixations increased in sleepy state (Figure 4). However, average fixation duration has noticeably dropped (Figure 5). Total time spent on fixations has shown no noticeable difference between sleep deprived and fresh states (Figure 6).

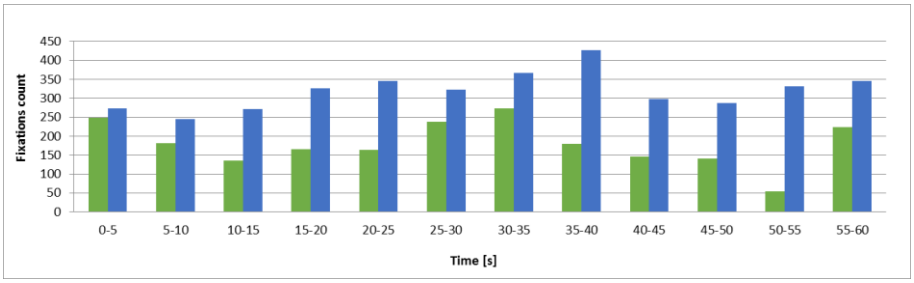


Figure 4. Fixation count analysis for one driver in fresh (green) and sleep deprived (blue) condition.

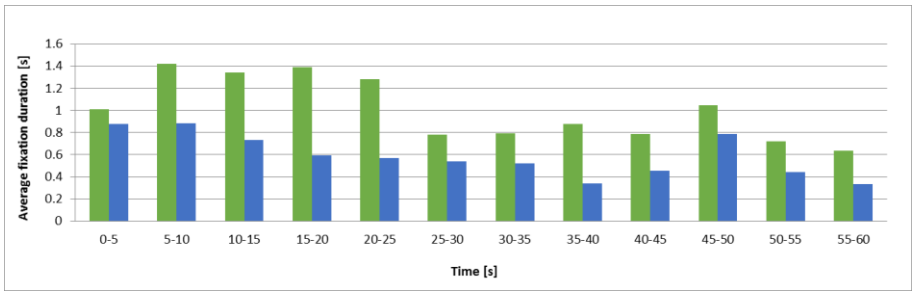


Figure 5. Average fixation duration analysis for one driver in fresh (green) and sleep deprived condition.

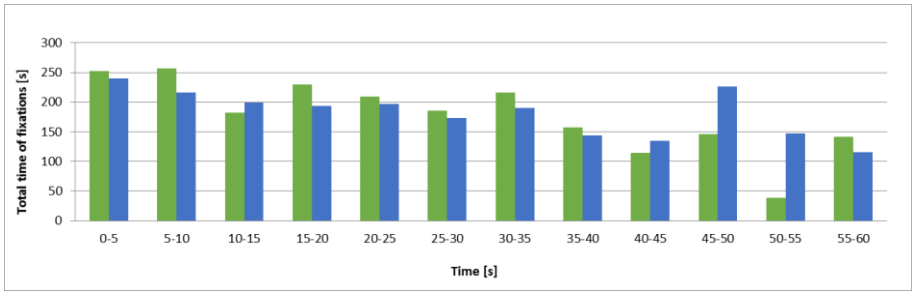


Figure 6. Total fixation duration analysis for one driver in fresh (green) and sleep deprived condition.

Blinks were also analysed for frequency, average and total duration during one hour of measurement (first hour in case of sleep deprived drivers) per 5-minute time sections. At this stage, no typical schemes of blink behaviour can be observed as behaviour of each participant is individual. In Figure 7, one can see that different tendencies of blink frequency are observed through one hour for one of the drivers (same participant, whose fixation analyses are suggested above): noticeable increase in blinks count for sleep deprived state. Blink measure provided in eye tracker data is an eyelid closure of duration under 700 ms. This could mean that in the sections with blinks count drop we may be facing sleep events. However, further analysis needs to be done for more solid assumptions. Total blinking time values per each time section are represented in Figure 9. In general, total blinking time over the whole period (60 minutes) hasn't changed between states as compared within 1 hour. Average blink duration is somewhat shorter for sleep deprived drivers (see Figure 8), however difference is not very representative, besides, the longer eye closures could be not detected due to blink measurement parameters described above. Further research will be concentrated at correlation of eye behaviour to driving behaviour measures as well as sleep events detection for better understanding of eventual behaviour changes observed here.

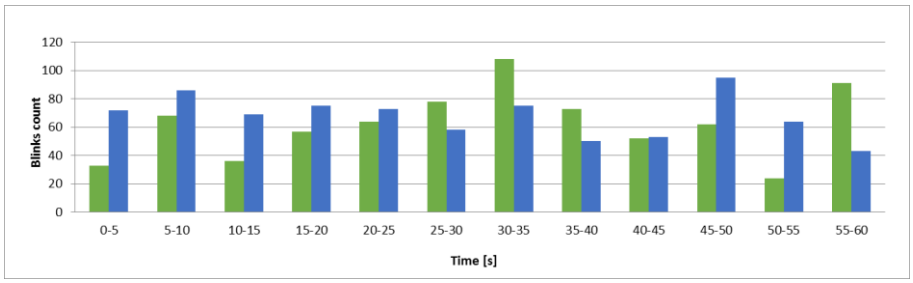


Figure 7. Blinks count analysis for one driver in fresh (green) and sleep deprived (blue) condition.

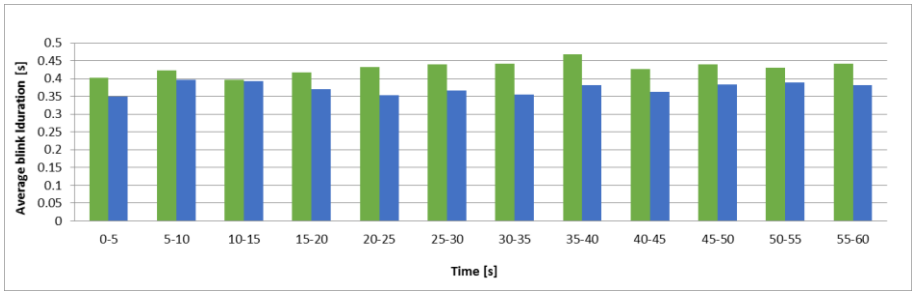


Figure 8. Average blink duration analysis for one driver in fresh (green) and sleep deprived (blue) condition.

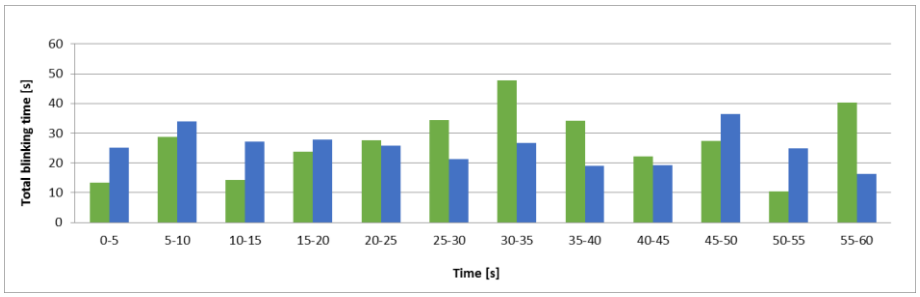


Figure 9. Total blinking time analysis for one driver in fresh (green) and sleep deprived (blue) condition.

Analysis of driving data

The incident method for experiments with sleepy drivers on a driving simulator is described by Baulk et al. (2006). The research shows that longer time spent on task by drivers after sleep deprivation was resulting in increase of reaction times.

The outcomes of experiment described here show that even though no significant differences have been noticed for drivers in fresh state (mean 1220 ms; SD 500 ms) as compared to sleepy state during the 1st hour (mean 1414 ms, SD 711 ms, t= 1.6, p=0.1) and 2nd hour (mean 1315; SD 447 ms, t=1.03, p=0.3), however, it is worth of noticing that the overall grow of time to react grew further in the course of the

sleepy experiment – the tendency can be tracked in Figure 10, where mean reactions among all subjects are displayed per each trigger event in fresh state, sleep deprived state (first and second hours of driving in sleep deprived state are displayed separately). For many drivers the abrupt breaking at trigger zones has been observed. Further research will be concentrated on more driving characteristics parameters.

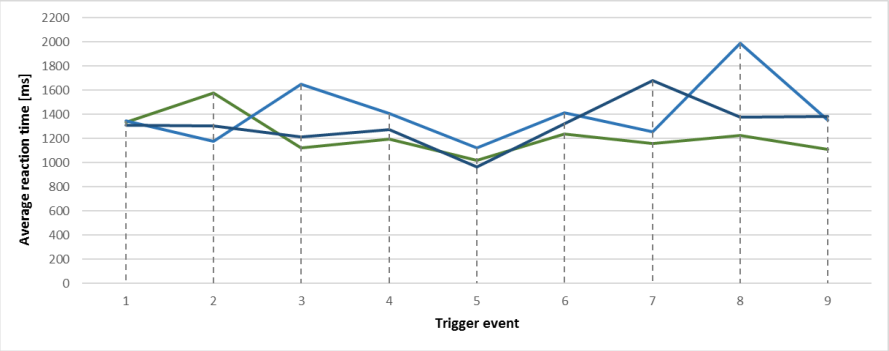


Figure 10. Average reaction time for each trigger event in each state: green - fresh, light blue - sleep deprived 1st hour, dark blue - sleep deprived 2nd hour.

Subjective measurements and some experiment findings

Normally people are aware of their sleepy state. However, studies show that they aren't capable of predicting crash or inability to drive (Watling et al., 2015; Horne & Baulk, 2003). It has been shown by Howard et al. (2014) and Åkerstedt et al. (2014) that self-rating increases with longer awake hours. Though not considered to be a sufficient measure of sleep prediction, self-rating is still worth of being taken into consideration especially when combined with objective evaluation techniques of sleepy driver behaviour. There are several sleepiness scales for subjective measure, such as Stanford Sleepiness Scale (Hoddes et al., 1973), Karolinska sleepiness scale (Kaida et al., 2006). Another self-rating scale has been developed at author's faculty, also mentioned by Bouchner et al. (2006), for driver self-assessing of their state. In this study, sleepy drivers were self-rating themselves during 18 stops designed as trigger events in simulation scenario. The scale itself is a result of researchers' observations during numerous experiments of driver fatigue. The outcomes of participants' self-ratings in the current study are presented in Tables 2 and 3.

Changes in subjective feeling of sleepiness are individual. The following tendencies have been noticed here: participants 3, 4 and 5 reported to be at the highest stages of going into sleep at the early stages of the experiment; at the same time, one participant reported to be in an average sleepy state. It is important to collect more data for detecting typical tendencies. Comparison to real driving behaviour by analysis of vehicle outputs is important here.

During after-experiment moderated conversations with drivers we were seeking to find possible personal perceptions on questions about feeling, mood, and physical state. Most drivers have either expressed a desire to use an air conditioner or radio,

some drivers tried to sing or read poems by heart. Such countermeasures were found by some researchers, such as, for example, Reyner and Horne (1998) and Schwarz and Ingre (2012) to be of little or zero effect on level of sleepiness. However, one might consider those as behavioural features of drowsy state for visual observations during driver sleepiness experiments.

Table 1. Sleepiness scale. Faculty of transportation Sciences CTU in Prague

Score	***State description as perceived by driver***
1	*I feel fine/fresh & driving does not make me any problems.*
2	*I feel drowsy & driving does not make me any problems.*
3	*I feel drowsy & I notice some problems.*
4	*I feel very drowsy & I need excessively concentrate to drive correctly.*
5	*I experienced 'blackouts' & losing of control over the car.*

Table 2. Self-rating scores provided by 5 participants during sleep deprived state.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
p1	2	3	3	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5
p2	1	2	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	5
p3	2	3	4	4	3	4	4	4	5	5	4	3	3	3	4	3	3	4
p4	2	4	5	5	5	5	4	5	5	4	5	5	5	5	5	5	5	5
p5	4	5	5	5	5	5	5	5	4	4	4	4	4	4	4	4	4	4

Note: Self-evaluation scores of 5 participants during 2 hours of driving (columns 1 to 9 correspond to the first hour, columns 10 to 18 correspond to the second hour). Scores from 1 to 5 are given per scale presented in Table 1. Light green cell colours (weak shading) represent lower score of drowsiness, blue colour (strong shading) represent higher score of drowsiness.

Discussion and future work

The versatile approach to driver drowsiness is explained by the different character of human behaviour in similar situations. Individual approach in human behaviour research shall be applied therefore. The provided research has analysed behaviour of sleepy drivers as compared to that in fresh state. The goal of the research is developing an efficient technique, based on visually available parameters, to detect a dangerous behaviour. The extensive study of broad data collection from the experiments includes objective data collected from simulator driving outputs to subjective evaluation of driver with the help of self-evaluation tests. Separate observations show some tendencies in deterioration of driver eye behaviour, determined by visual behaviour changes, such as decrease in average duration of gaze fixation count and increase of fixation count. It was also possible to observe

increase of blinking frequency, however blink length average values didn't display noticeable difference between two compared states. The results differ per individual and solid conclusions require measurement of bigger cohort. Calibration of analysis of eye behaviour is needed for specification particular conditions of certain behaviour and for better correlation of visual behaviour to objective data results and subjective evaluation. Each type of obtained results is depending on different parameters, such as time, i.e. time from last sleep, physical condition of each individual, driving experience, collection of quality visual data is complicated by various structure and in-cabin behaviour habits (body movements, gestures), not to mention the aspects of sleepy driver face relaxation and eyes half-closures. Obtained objective measures showed insignificant statistical differences. However, separate critical values of reaction times were noticed in several cases. Further cross analysis of experimental data and research shall continue.

Acknowledgement

This research was supported by grant SGS16/254/OHK2/3T/16 "Experimental research of driver fatigue by means of observation visual behavior".

References

Akerstedt, T., Anund A., Axelsson J., & Kecklund G. (2014). Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *J Sleep Research 23*, 240-252.

Anund, A. (2009). *Sleepiness at the wheel*. PhD thesis, Karolinska Institute. Stockholm, Sweden: NASP – National Prevention of Suicide and Mental Health.

Baulk, S., Biggs, S., Heuvel, C., Reid, K., & Dawson, D. (2006). *Managing driver fatigue: quantifying real world performance impairment*. (ATSB Research and analysis report). Australian Transport Safety Bureau.

Bergstrand, M. (2008). Automatic analysis of eye tracker data. Swedish National Road and Transport Research Institute.

Bouchner, P., Novotný, S., Hajný, M., Piekník, R., Pěkný, J. & Valtrová, K. (2006). Analysis of technical and biological outputs from simulated driving, focused on driver's fatigue detection. DSC Asia/Pacific, Tsukuba.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R. & Dement, W. C. (1973), Quantification of Sleepiness: A New Approach. *Psychophysiology, 10,* 431–436.

Hooge, I.Th.C. & Erkelens, C.J. (1997). Adjustment of fixation duration in visual search. *Vision Research, 38,* 1295-1302.

Horne, J.A. & Baulk, S.D. (2003). Awareness of sleepiness when driving. *Psychophysiology, 41*, 161-165.

Howard, M.E., Jackson, M.L., Berlowitz, D., Fergal, O., Swann, P., Westlake, J., Wilkinson, V. & Pierce R.J. (2014). Specific sleepiness symptoms are indicators of performance impairment during sleep deprivation. *Accident Analysis and Prevention 62*, 1-8.

ISO 2014, Road Vehicles – Measurement of driver visual behavior with respect to transport information and control systems – Part 1: Definitions and parameters ISO 15007-1.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. (Psychological review V87/No. 4). Carnegie Mellon University. Department of Psychology.

Kaida, K., Takahashi M., skerstedt T., Nakata A., Otsuka Y., Haratani T., Fukasawa K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. Clinical Neurophysiology 117 (pp. 1574-1581).

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin, 124*, 372-422.

Reyner, L.A., Horne, J.A., 1998a. Evaluation "in-car" countermeasures to sleepiness: cold air and radio. Sleep 21 (1), 46–50.

Schwarz, J.F., Ingre, M., Fors, C., Anund, A., Kecklund, G., Taillard, J., Åkerstedt, T. (2012). In-car countermeasures open window and music revisited on the real road: popular but hardly effective against driver sleepiness. *Journal of Sleep Research, 21*, 595–599.

Watling C.N., Armstrong K., Radun I. (2015). Examining signs of driver sleepiness, usage of sleepiness countermeasures and the associations with sleepy driving behaviours and individual factors. *Accident Analysis and Prevention, 85*, 22-29.

Can User Experience affect buying intention? A case study on the evaluation of exercise equipment

Giuseppe Fedele[1], Mario Fedriga[1], Silvano Zanuso[1], Simon Mastrangelo[2], & Francesco Di Nocera[3]
[1]Scientific Research Department, Technogym S.P.A., Cesena, Italy - [2]Ergoproject S.r.l., Rome, Italy - [3]Sapienza University of Rome, Italy

Abstract

Treadmills are increasingly loaded with digital technology for assisting the individual during the workout sessions by providing information for tracking relevant training parameters. Also, this technology makes exercise more pleasurable by keeping the user connected to her/his digital ecosystem (e.g. social networking, access to multimedia content). Although there is an increasing interest in digital technologies to be used in fitness, a cursory literature search shows that the interest towards gym equipment is currently limited to the hardware component, thus making biomechanics the focus of the investigation. Other types of contributions are very rare and mostly focused on the design of tools for special populations (e.g. elderly, disabilities) as well as for promoting physical activity monitoring (eHealth). In the present study information on the perceived usability of the interface was collected and analysed along with opinions about buying intention and estimated pricing. Twenty-three individuals were tested after using a treadmill (Technogym S.p.A.) equipped with an interface allowing equipment and training management, activity monitoring and user entertainment. Results indicated a significant influence of perceived usability of the interface on the intention of buying the whole system, thus suggesting the existence of a ROI of Human-Centred Design strategies.

Introduction

The growing interest towards Usability and User Experience (UX) is not limited to the goal of devising better design strategies, but it is also related to the increasing awareness about the relation of these aspects with the internal (e.g. staff productivity, software development costs) and external (e.g. conversions, buying intention) Return On Investment (ROI), that is the benefit to an investor (e.g. sales increase, enhanced brand perception) resulting from an investment of some resource (e.g. internal effort, consulting).

Although the commercial impact of usability has been acknowledged since the seventies (Bennet, 1979), and has been elaborated over the years giving rise to the multifaceted UX construct (e.g. Bias & Mayhew, 2005; Graefe, Keenan & Bowen, 2003; Nielsen et al, 2008; Watermark Consulting, 2015), the real understanding of

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

the causal (interaction) variables influencing the ROI is still unclear (Rosenberg, 2004; Weinschenk, 2005). Factors influencing this lack of knowledge may be the reluctance of business owners in disclosing the details of their success/failures, as well as the fact that the ROI needs to be related to multiple aspects, such as:

- UX activities frequency and typology;
- type of product;
- specific sector in which interfaces are implemented.

The recent trends (e.g. Internet of Things) have further complicated the scenario considering that nowadays digital interfaces (not necessarily GUIs) are embedded in any kind of product (e.g. cars, home appliances, wearables) and it becomes more and more complicated to understand which part of the product could have an higher impact on ROI (e.g. design, comfort, usability).

The gym is one of those complex UX emerging markets. The so-called "smart gym" is a new paradigm aimed at supporting both users and trainers in keeping track of the work-out activity and tailor it to the real needs and capacity of the individual. Jain (2015) applied this approach to two different equipment pieces (squat machine and leg press machine) matching the work-out activity to the individual profile in order to automatically calibrate and modify the exercise schedule and the workload. The system has been designed to achieve the following three objectives: 1) to provide user with personalized system-generated workout suggestions; 2) to track the user activities and maintain individual history records; 3) adjust the workout regimen according to the available resources. They compared this approach to traditional workout sessions with a personal trainer and found that the adaptive system provided better results in terms of displacement, force and time elapsed. As a matter of fact, gym equipment has been increasingly loaded with digital technology for assisting the individual during the workout sessions, as well as for making the physical activity more pleasurable. In many cases touchscreen displays not only provide information (and require input) that is relevant for the exercise (e.g. providing information about the heart rate) but also for keeping the users connected to their digital ecosystem (e.g. social networking, personalized multimedia content). That makes the usability of these tools particularly interesting to the HF/E community. Unfortunately, although there is an increasing interest towards mobile technologies to be used in fitness, a cursory literature search can easily show that the interest towards gym technology (particularly, treadmills) is currently mainly limited to the hardware component, thus making biomechanics the focus of the ergonomics investigation with some reference to comfort and safety (Biscarini, 2012; Carraro et al., 2014; Reilly & Lees, 1984). Other types of contributions are very rare, and mostly focused on the design of tools for special populations (e.g. elderly, people with disabilities) as well as for promoting physical activity (the so-called “exergames”: see Mueller et al., 2011) and monitoring (the so-called “eHealth”).

Moreover, there is a complete lack of studies trying to investigate the relation between UX and ROI. The case study reported here is a first attempt to deal with this topic in the context of fitness equipment. Particularly, we asked a selected group of individuals to perform a workout session using a treadmill equipped with a companion Graphical User Interface (GUI) and then to evaluate their experience

with the system and their willingness to promote and pay. It is expected to find a link between the experience with the system and the willingness to pay that, in turn, would suggest that investments in UX are as profitable as those in other aspects of both the design and the marketing of a product.

Study

Participants

Twenty-three individuals (12 males, 11 females) volunteered in this study. They belonged to 4 age groups: 25-35 years old (N=8); 35-45 years old (N=8); 45-55 years old (N=4); 55-65 years old (N=3). Structured questionnaires were used to collect information regarding type and intensity of usually practiced physical activity and regarding the familiarity with mobile applications (apps). Generally, the selected user profile included users who often attend the gym and employ the treadmill as the main exercise equipment. Five participants reported to never use mobile apps.

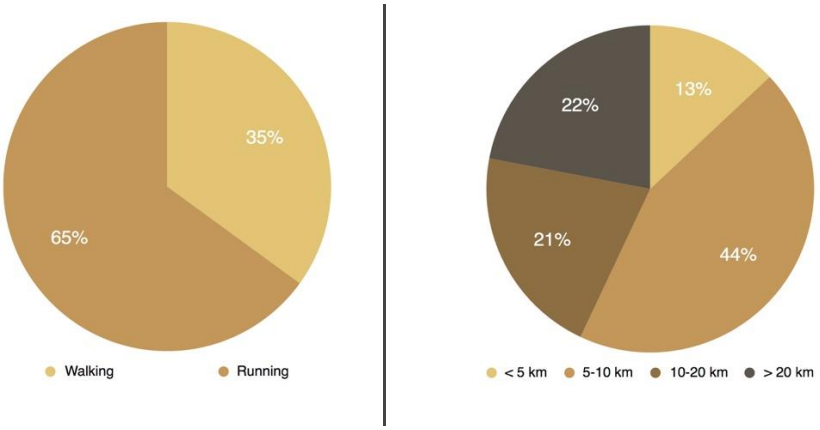


Figure 1. Distribution by favorite workout activity (walking vs. running, left pane) and average km per week (right pane).

Equipment and setup

Product description

A consumer Technogym treadmill (for domestic use only) was used in this study. The treadmill can be used by itself or connected to a native app on a tablet. The treadmill features four indicators (running time, machine inclination, running speed, covered distance) and three physical buttons for starting / stopping the running session and for controlling machine inclination and running speed. Above the machine it is possible to insert a tablet for using the MyRun app (available both for iOS and Android systems) through which the users can create a custom program based on their physical fitness and the frequency of workouts. The app automatically selects music that is matched to users' pace (e.g. through a Spotify account) and compares running sessions on the treadmill with outdoor performance tracked by

compatible apps (e.g. Runkeeper, Mapmyfitness) or specific tracking devices (i.e. Fitbit and Polar).

Figure 2. Technogym treadmill (left pane) equipped with a tablet running the MyRun App (right pane).

Performance metrics

Performance was measured as the proportion of errors made while completing the following tasks (manually recorded by one of the authors):

1. Setting the training goal (i.e. starting a running session of a predetermined duration and keeping a presetted speed).
2. Starting / stopping the training session (either using the app or the physical button placed on the treadmill).

Subjective metrics

Opinions/judgment expressed by users were collected at the end of each session and provided information about three diffecter constructs using the following tools:

- User Experience (UX): participants were asked to answer to four True / False questions (see table 1) related to four dimensions (distractibility, ease of use and pleasantness, functionality).
- Mental Workload: participants rated their experience using the NASA-TLX (Hart & Staveland, 1988).
- Promotion: participants rated their willingness to promote the equipment using the Net Promoter Score® (NPS: Reichheld, 2003), a metric that provides an estimate of the probability with which a person is willing to recommend a particular product to other people on a scale from 0 to 10.
- Willingness to Pay (WTP): participants reported of the maximum price they were willing to pay on the purchase of the same product. They also were requested to provide an estimate (guess) of the actual price of the product.

Additionally, anagraphic data and information on exercise preference (walking vs. running) was collected at the beginning of the session.

Table 1 – UX "quick and dirty" scale. Total score was obtained by summing up all "true" answers after reversing item 1 and 4.

1.	The app distracts me from those aspects I normally pay attention to (e.g. distance travelled, calories burned).	TRUE	FALSE
2.	The app is easy to use.	TRUE	FALSE
3.	The app is pleasant and well finished.	TRUE	FALSE
4.	The app did NOT help me during the workout.	TRUE	FALSE

Procedure

Participants were greeted and introduced to the tasks to carry out using the treadmill. They were asked to sign an informed consent and to fill a personal data form and then to get on the treadmill, set a 10 minutes session and start running and/or walking (depending on their exercise habits) at a speed of 5 km/h at least, while interacting with the MyRun app. During the session they were free to express what they thought about the product, but were not encouraged to do so as in a thinking aloud protocol. The questionnaires described in previous section were administered at the end of the interaction with the product and a debriefing section concluded the session.

Data analysis and results

Answers to the four items of the questionnaire on the quality of interaction with the system were summed up and participants were classified as showing Negative UX (N=8) if their score was below the median. All other participant were classified as showing Positive UX (N=15). The proportion of errors on the number tasks to be accomplished (choose a goal-based workout session, set the time as the goal of the session, set the duration of the session, start the session, stop the session) was used as dependent variable in two ANOVA designs Exercise Preference (Walking vs. Running) x Gender (Males vs. Females) and UX (Positive vs. Negative) x Gender (Males vs. Females), respectively. Results showed no significant differences for both analyses (exercise preference: $F_{1,19}$=1.15, p>.05; gender: $F_{1,19}$=.51, p>.05; interaction: $F_{1,19}$=2.02, p>.05 and UX: $F_{1,19}$=.22, p>.05; gender: $F_{1,19}$=.04, p>.05; interacton: $F_{1,19}$=.09, p>.05, respectively).

NASA-TLX score was used as dependent variable in two ANOVA designs Exercise Preference (Walking vs. Running) x Gender (Males vs. Females) and UX (Positive vs. Negative) x Gender (Males vs. Females), respectively. Results of the first analysis showed no significant effects (exercise preference: $F_{1,19}$=1.16, p>.05; gender: $F_{1,19}$=.06, p>.05; interaction: $F_{1,19}$=1,67, p>.05).

A significant main effect of UX ($F_{1,19}$=4.30, p=.05) was found in the second analysis: users who reported a positive UX also reported lower mental workload. Neither a significant effect of gender ($F_{1,19}$=3.08, p>.05) nor a UX by gender interaction ($F_{1,19}$=.75, p>.05) was found.

Net Promoter Score (NPS raw score) was used as dependent variable in an ANOVA design Gender (Males vs. Females) x UX (Positive vs. Negative). Results showed a significant main effect of UX ($F_{1,19}$=11.59, p>.01): users who reported a positive UX

also reported higher NPS values. Neither a significant effect of gender ($F_{1,19}$=.15, p>.05) nor a UX by gender interaction ($F_{1,19}$=.06, p>.05) was found.

Willingness to Pay (WTP), that is the maximum price at or below which a consumer will definitely buy one unit of the product, was used as dependent variable in an ANCOVA design using UX (Positive vs. Negative) as factor and Estimated Price as covariate, which was found to be positively correlated to WTP (r=.49; p<.05). The covariate was introduced for subtracting the weight of the estimated (guessed) value of the product. Results showed a significant effect of the covariate ($F_{1,20}$=6.57, p<.05) and a tendency toward statistical significance of UX (F1,20=3.92, p=.06): users who reported a positive UX also reported higher WTP close to the real price of the product.

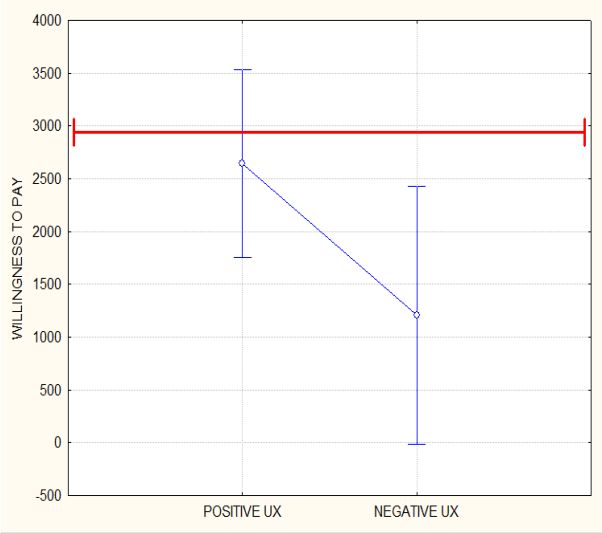


Figure 3. Willingness To Pay (WTP) in Euros by User Experience. Vertical bars denote .95 confidence intervals and the red line indicates the actual price of the equipment tested. Notably, the mean WTP for the Positive UX sub-group Approached the actual price of the equipment.

Discussion and conclusion

This case study showed that UX with the MyRun app influenced the attitude towards the entire product. All participants showed homogeneous performance in terms of number of tasks correctly accomplished during the experimental workout session, but users whose experience with the treadmill was positive reported lower mental workload compared to those who experienced a negative interaction with the system. The relation between usability and workload is often overlooked. Nevertheless, problems related to usability may lead to a loss of control, orientation and frustration from the user (e.g. Faliagka, Rigou & Sirmakessis, 2015) and -as reported by Longo (2015)- “mental workload is gaining momentum as an important

design concept in human–computer interaction and is important in considering the interaction of people with computers and other technological devices" (p. 758).

Satisfied users reported higher NPS values, therefore indicating them as advocates and promoters of the product, and also reported higher willingness to pay for the equipment a price that is close to the actual value of the good. In other fields, such as e-commerce, the influence of usability on buyers' intention has been already reported (Konradt et al., 2003), and this result is aligned with that literature.

Interestingly, even if the estimated price resulted to be positively correlated to the willingness to pay, the effect was obtained also after subtracting the influence of the estimated price: no matter what is the imagined monetary value of the product, individuals who experience a better interaction with the system are willing to pay for it and, very surprisingly, they are willing to pay its actual price (even if they has no information about the actual price).

So far, the influence of UX on promotion was reported only for web sites in correlational studies (Sauro, 2012). The present case study is the first to take into account this relation for physical products and confirms what has been found elsewhere: the usability of a system is a strong determinant of success for a product / service. The ROI of UX research and consultancy is a fact and should be taken into consideration by any organisation / company.

References

Bennett, J. (1979). The commercial impact of usability in interactive systems. In: Shackel, B. (ED.), *Man/computer Communication: Infotech State of the Art Report, vol. 2* (pp. 1-17). Maidenhead, UK: Infotech International.

Bias, R.G., & Mayhew, D.J. (Eds.). (2005). *Cost-justifying usability: An update for the Internet age*. Burlington (MA): Elsevier.

Biscarini, A. (2012). Measurement of power in selectorized strength-training equipment. *Journal of Applied Biomechanics, 28*, 229-241.

Carraro, A., Gobbi, E., Ferri, I., Benvenuti, P., & Zanuso, S. (2014). Enjoyment perception during exercise with aerobic machines. *Perceptual & Motor Skills, 119*, 146-155.

Faliagka, E., Rigou, M., & Sirmakessis, S. (2015). A usability study of iPhone built-in applications. *Behaviour & Information Technology, 34*, 799-808.

Graefe, T.M., Keenan, S.L., & Bowen, K.C. (2003). Meeting the challenge of measuring return on investment for user centered development. *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, 860-861.

Konradt, U., Wandke, H., Balazs, B. & Christophersen, T. (2003). Usability in online shops: scale construction, validation and the influence on the buyers' intention and decision. *Behaviour & Information Technology, 22*, 165-174.

Jain, A. (2015). A Smart Gym Framework: Theoretical Approach. In *Proceedings of the IEEE International Symposium on Nanoelectronic and Information Systems*, 191-196.

Longo, L. (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology, 34*, 758-786.

Mueller, F.F., Edge, D., Vetere, F., Gibbs, M.R., Agamanolis, S., Bongers, B., & Sheridan, J.G. (2011). Designing sports: a framework for exertion games. *Proceedings of ACM CHI*, 2651-2660.

Nielsen, J., Berger, J., Gilutz, S., & Whitenton, K. (2008). *Return on Investment (ROI) for usability* (4th ed). Freemont, CA: Nielsen Norman Group.

Reichheld, F.F. (2003). The One Number You Need to Grow. *Harvard Business Review, 81*, 46-54.

Reilly, T., & Lees, A. (1984). Exercise and sports equipment: some ergonomics aspects. *Applied Ergonomics, 15*, 259-279.

Rosenberg, D. (2004). The myths of usability ROI. *ACM Interactions, 11*, 22-29.

Watermark Consulting (2015). 2015 Customer Experience ROI Study ROI Study. Retireved from http://www.watermarkconsult.net

Weinschenk, S. (2005). *Usability: A Business Case* (white paper). Fairfield, IA: Human Factors International.

From aircraft to e-government - using NASA-TLX to study the digital native's enrolment experience for a compulsory e-service

Chris Porter
Faculty of ICT, University of Malta
Malta

In recent years Malta launched a new e-service for students aged 16–18 who are applying for national exams. Adoption is compulsory and students also need to enrol for a national e-ID to gain access to the service. The e-service enrolment process is a pivotal part of the user experience, and without proper considerations it can become a major hurdle, stopping users from transacting online. This paper presents results from a two-stage study conducted with affected students to (1) measure and assess the impact of enrolment-specific design decisions on the students' lived experience (using NASA-TLX as a multi-dimensional and subjective workload assessment technique) and to (2) validate and critically assesses NASA-TLX's applicability and sensitivity in this context. This study gives particular attention to digital natives – people who have grown up with and are highly accustomed to digital technology (Prensky, 2001). This study shows that NASA-TLX is reasonably sensitive to changes in workload arising from various design-decisions within this context, however certain adoption caveats exist: (1) unsupervised NASA-TLX participants may provide significantly different results from supervised participants for most workload scales, (2) context-specific definitions and examples are necessary for most workload scales and (3) there are no major advantages arising from the adoption of a mean weighted workload (MWW) metric over raw TLX (RTLX).

Introduction and Aims

The enrolment process for any e-service can have a significant impact on the user's lived experience (Porter et al., 2012) and in turn on the success of the e-government service itself (Axelsson & Melin, 2012). In Western Europe the first generation of digital natives are starting to use e-government services. Most of these services require an online identity, which first-time e-government users have to create. The aims of this paper are to (1) develop an understanding on how enrolment-specific workload, as a multidimensional measure, impacts the digital native's experience with online services and (2) whether NASA-TLX is a suitable candidate to, in part, answer this question. Qualitative techniques are used to capture this citizen group's perceptions, expectations and reactions to identity-related tasks. This study also aims to determine whether NASA-TLX (1) is easy to understand and follow for younger (and untrained) participants, (2) whether it is applicable within this particular context (e-government) and (3) whether it is sensitive enough to detect changes in workload arising from different e-service enrolment process designs. Qualitative

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

results will be treated as out of scope for this paper, and will not be presented nor discussed.

Background

Hart describes workload as 'the cost of accomplishing mission requirements for the human operator' (2006, par 1). The human costs in e-government include the citizen's inability to use an e-service - which could in turn result in sanctions, such as a fine for not paying a congestion charge on time (Inglesant & Sasse, 2007), or loss of opportunities, such as having to use otherwise productive time to visit a government department in person. However, the risk is also on the service owner: if the human cost for a service is such that e-services are not used, the government will also have to absorb costs for handling that particular transaction via traditional channels. This can also have political ramifications in that a negative experience will generally reflect negatively upon the government's image of efficiency and competence.

Cain argues that different workload measurement techniques actually assess different aspects of workload and this heterogeneity of focus stems from the 'lack of an accepted definition of workload' (2007, pg 7). According to the author different people have different perspectives on the meaning of workload, including (1) the task demands imposed on the user, (2) the effort the user needs to make to satisfy such demands and (3) the consequences of attempting to meet such demands. Sasse et al. (2014) adopted the GOMS-KLM approach (Goals, Operators, Methods and Selection rules – Keystroke-level Modelling) to assess the workload imposed by authentication events in terms of the time taken for a user to complete them. GOMS-KLM, introduced by Card et al. (1980), evaluates workload by deconstructing tasks into a set of basic actions, or steps, on which time measurements are taken. Although it is an important benchmarking technique to help practitioners determine the best and worst case scenarios in terms of user performance and effort for a given task, its simplicity might deter from its potential to provide measurable information on aspects such as frustration and self-confidence which, from a lived experience perspective, are also important considerations for the design of better security mechanisms. The author believes that the time taken to fill in a form does not necessarily imply a negative user experience, especially if the benefits obtained from using the e-service offset the cost associated with accessing it. For instance, a tax return e-service requiring users to authenticate by selecting a digital certificate, submitting a one-time password and filling in several other fields might still be worth the while for a professional who would otherwise need to regularly fill in and post paper-based forms on behalf of his clients. Workload can affect users in different ways, and for different reasons, and this impact may also vary across contexts of use.

For this reason, the author turned his attention to NASA-TLX – a multi-dimensional and subjective workload assessment technique. While developing NASA-TLX, Hart and Staveland (1988) examined ten workload-related factors, retrieved from sixteen experiments. Six of these factors were then proposed as a multi-dimensional rating scale combining magnitude and source information 'to derive a sensitive and reliable estimate of workload' (Hart and Staveland, 1988, pg 139). This was accomplished

after a series of statistical tasks, mainly to determine the sensitivity of each factor on workload. In NASA-TLX, physical and mental workload are also measured along with cognitive workload. This technique was originally developed for use in aviation flight-deck design; however, now it has been widely adopted for alternative uses and is also being used as a benchmark against which other workload measuring techniques are evaluated. Rubio et al. (2004) surveyed a number of studies which adopted subjective workload (cognitive) rating techniques. The authors ranked NASA-TLX at the forefront of sensitivity to experimental changes in workload conditions. This is also confirmed in Garteur's *Handbook of Mental Workload Measurement* (Garteur, 2003). Hill et al. (1992) also rated NASA-TLX as the most sensitive to workload changes, followed by MCH (Modified Cooper-Harper) and finally SWAT (Subjective Workload Assessment Technique).

NASA-TLX allows subjects to record data post-task, and thus certain physiological and time span-dependent effects may be in conflict to what is recalled by the subject. Techniques to counteract this issue include (1) screen-recording playback and (2) video-recording playback of the tasks performed. These techniques are designed to facilitate retrospective workload rating (Garteur, 2003). NASA-TLX uses six workload factors, or dimensions, and measures their relative contribution in influencing the user's perceived overall workload. Twenty years after presenting NASA-TLX, Hart (2006) reviewed the current state of the technique. It was found that most recent studies using this technique handled investigations on interface design and evaluation, with 31% focusing on visual and auditory displays and 11% on input devices. Seven percent of the studies were carried out with users of personal computers. Hart notes that NASA-TLX can be used in various situations, from aircraft certification to website design. This study proposes the use of NASA-TLX to measure enrolment-specific workload, primarily because of its multi-dimensional nature and overall performance sensitivity. Various other advantages of NASA-TLX include: ease of use; practicality of the method; reduction of between-rater variability (due to the adoption of weighted rankings) and the availability of clear instructions, supporting tools and case studies.

Study Context

The examinations department stipulated that students are to use a new e-service to register for their A-level examinations. Unless there were exceptional circumstances, students could not apply via the traditional method of visiting the examinations department in person. A 'Click Here to Apply' button was made available on a clean and easy to follow landing page at https://exams.gov.mt. Once clicked, students were asked to login using their e-ID credentials. No immediate information is given on how to obtain an e-ID. Instructions on how to enrol for an e-ID were provided in another e-government page, and at the time the process consisted of the following steps:

1. Visit the registration office in person (on average it takes 30 minutes each way by bus)
2. Go through a short enrolment process (on average it takes 5 minutes to complete and students need to present their national ID card and a valid

email address). Queues are possible since this is a central-government office
3. Receive a security PIN by post at the address given at enrolment
4. Activate the e-ID account using the PIN received by post and a password received at the email provided in step 2
5. Create a new password that adheres to a strict password policy

Once students are successfully enrolled on the National Identity Register, they are able to proceed to register and pay for their A-level examinations through the e-service website at https://exams.gov.mt/.

Method

Participants

Two sets of participants were involved in this study, one for each of the two phases discussed below, namely the (1) collection and analysis of users' experiences via an online questionnaire and the (2) follow-up workshops to verify and validate NASA-TLX ratings.

Process

The author's goal was to capture as much feedback as possible from the pool of students sitting for their exams. An online questionnaire was opted for since it would help (1) reach as many students as possible while (2) minimising disruptions to their studies. A number of interesting insights and recommendations emerged during this exercise. It was also felt that this study would benefit highly from a second intervention through which the initial results could be validated and substantiated. This was the motivation for the second part of the study which offered the opportunity to assess the applicability and understandability of NASA-TLX with digital natives and to investigate its sensitivity towards workload induced by enrolment-specific factors. Students who indicated that they would be willing to participate in follow-up meetings were contacted and a series of five workshops were scheduled. All ethical considerations recommended by the research ethics committees at the respective institutions were observed for both phases of the study.

Results

Three data sets were generated following this study: (1) qualitative results from the questionnaire outlining experiences for the various subgroups, (2) quantitative workload data obtained from the questionnaire's NASA-TLX assessment and (3) data from follow-up sessions which includes both qualitative and TLX related information. Thematic insights arising from the transcribed qualitative data will not be presented here.

Unsupervised NASA-TLX – online questionnaires

The questionnaire was sent to over 1000 students who were sitting for their A-Level examination sessions. A total of 134 valid responses were received (13% response

rate). Sixty-two percent of the participants were female, 21% male and 17% decided not to disclose their gender. Eighty-one percent of students declared that they fall within the 16–18 age-group, while 15% chose not to disclose their age. Four participants stated that they are aged 19-24, and one was over 25 years of age. Only those falling within the 16–18 age bracket were considered in the analysis stage. Furthermore, around 10% (13) of the respondents accepted the invitation to participate in one of a series of follow-up workshops held in the following months.

The second part of the questionnaire was an online version of the TLX workload assessment procedure. Initially students were asked to rate the six sub-scales (or workload dimensions) for the exam registration task (including e-ID enrolment if applicable), followed by the pairwise comparison to get a weighted overall workload measure (mean of weighted ratings). The six workload dimensions are Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Own Performance (P), Effort (E) and Frustration (F). The overall task load index (MWW) was calculated for each participant, and averaged across the various student subgroups (see Figure 1). The cohort who used the e-service to complete the task, provided an overall mean weighted workload (MWW) of 42 (± 18.59) while those who registered for their exams in-person reported an overall MWW of 57 (± 14.11). These values, and particularly the variance in the online task's MWW, are not enough to draw any reliable conclusions on the users' experience. It would therefore be necessary to drill down into the various sources of workload while also analysing the process through which these values have been produced.

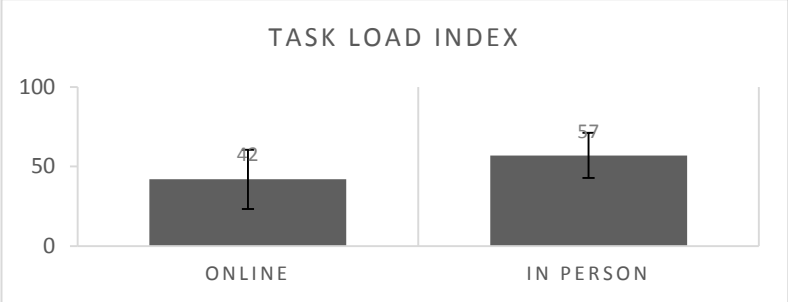


Figure 1. Mean weighted workload (MWW) for e-service users (online) and for those who adopted the offline exam registration process (at the exams registration department).

The average rating for the online method takes into consideration the ratings given by students who already had an e-ID and also by those who had to enrol for one. Table 1 shows how students who already owned an e-ID weighted the different workload dimensions.

Table 1. Workload dimension weighting by students who used the e-service and who already owned an e-ID

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	3.7	0.4	1.9	2.6	2.2	4.3
Median	4	0	2	3	2	4
Std.Dev	1.2	0.9	1.1	1.4	0.9	0.7

Table 2. Workload dimension weighting by students who used the e-service but had to enroll for an e-ID

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	2.8	1.4	3	2.1	2.2	3.7
Median	3	1	3	2	2	4
Std.Dev	1.4	1.5	1.4	1.4	1.1	1.3

Adjusted ratings are obtained by combining these weighted dimension values with raw ratings, as shown in Figure 2. In this case, Physical Demand is the lowest contributor to workload (adjusted rating = 2.5) however Frustration has an adjusted rating of 247, making it the highest contributor. Mental Demand follows Frustration, and thus these have a great influence on the average overall MWW. On the other hand, Table 2 shows how students who had to enrol for an e-ID weighted the different workload dimensions (out of 5). Figure 3 shows the respective adjusted ratings for this group. At a glance it is evident that this group of students had a different experience than the previous group and reported an increase in Physical and Temporal Demand. Frustration is still the highest contributor to workload, given an average weighting of 3.7, followed by Temporal Demand (3).

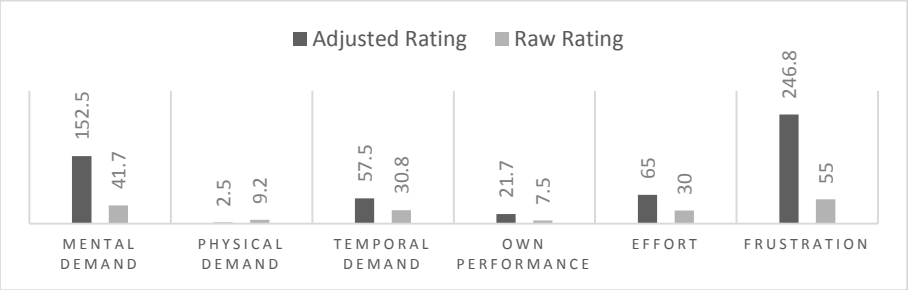


Figure 2. Adjusted rating for e-service users who already owned an e-ID (adjust rating = workload dimension weighting x raw rating)

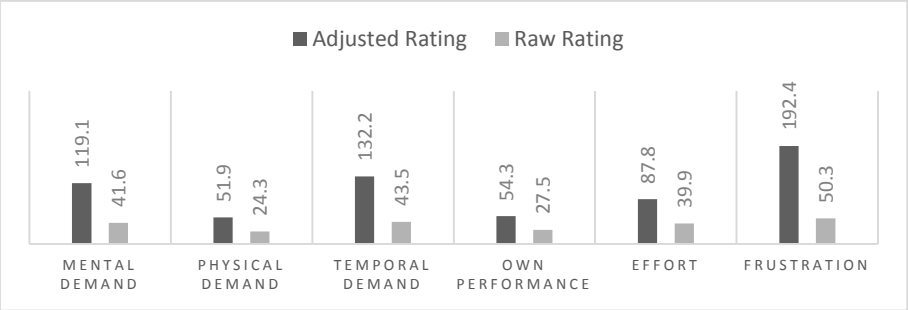


Figure 3. Adjusted rating for e-service users who had to enroll for an e-ID (adjust rating = workload dimension weight x raw rating)

Given this information, it can be seen that both groups of students (those who already had an e-ID and those who had to enrol for one) exhibited high levels of workload, albeit, for different reasons:

- *Those who had an e-ID:* Overall Task Load Index (TLX) was high mainly due to Frustration and Mental Demand. Causes for this outcome were various, including lack of process clarity, preference for traditional means, lack of trust in online systems and site performance.
- *Those who did not have an e-ID:* Overall TLX was high due to Frustration, Temporal and Mental Demand. Causes for this outcome were various, mostly due to the hassle involved to get an e-ID (e.g., waiting time at the e-ID enrolment office). Physical Demand was also significantly higher than that reported by the previous subgroup.

Supervised NASA-TLX – follow-up workshops

Students who agreed to participate in follow-up sessions were first asked to discuss their experience with the exam registration process and compulsory e-ID enrolment. Following this they were asked to compare and rate the perceived effort required to enrol for various online services including social networks, e-learning tools, payment gateways, email services and e-commerce sites. Each group had to reach a consensus for each rating decision and their interaction was observed. Following this, students were asked to go through a set of nine fictitious enrolment processes upon which workload measurements were taken. In all, 13 students agreed to participate in a series of follow-up sessions in small groups, eight of whom were female and five of whom were male. Their median age was 17 years old. All participants had just finished their A-level examinations

Perceptions on workload for popular online services

Before delving into the supervised NASA-TLX exercise, it was decided to conduct a series of semi-structured group-discussions without the use of rigid workload measurement techniques. This allowed for a consensus-driven thought process on the concept of workload as well as merits and de-merits of different enrolment processes adopted in popular online services. Each group of students (of 2 to 4 participants) was presented with a list of online services that they might have used at any point in time (e.g., Gmail, Facebook, Skype, PayPal and Hotmail amongst others). The most commonly used services for each group were then listed on a white board next to a rating scale indicating the level of perceived effort required to enrol for that specific service (i.e., easy, medium, difficult/annoying).

Figure 4. Participants had to agree on the level of perceived enrolment-specific workload (from personal experience) for several online services

First, students discussed the elements in enrolment they thought were responsible for workload from an individual perspective (this data will not be presented in this paper as it is deemed to be out of scope). Furthermore an agreement had to be reached on the relative level of perceived workload for each services' enrolment process in relation to other services' (as a group). Both mean and median values for the most commonly used services across all groups are presented in Figure 5. Feedback provided by different groups was normalised according to each group's rating patterns; that is, some groups always rated high, while others were more conservative. This made it possible to generate high-level, cross-group observations. Table 3 adds some context to these scores, providing annotations for the respective services' enrolment processes.

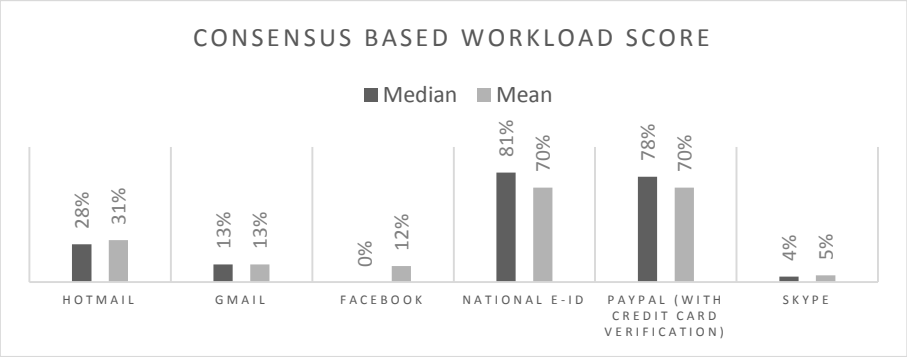


Figure 5. Perceived enrolment-specific workload for the most common online services

In a previous qualitative investigation (see Porter et al., 2013), it was established that; Items to Recall (ItR), Items to Generate (ItG), Interruptions to daily routines (I) and Delays (D) are central themes when it comes to sources of workload within enrolment processes. ItR represents the number of fields a user has to fill during the enrolment process. ItG represents a measure of the number of secrets the user has to

come up with (e.g., PIN, password). Any major interruption necessitating the user to go out of her way to complete the task is represented through I (e.g., visit an office to complete the process). Finally, D represents any form of interruption which introduces a delay in the process itself, but without disrupting the users' daily routines (e.g., a minor delay is introduced when an activation email is sent to the user, whereas a major delay is introduced when the service provider requires a day or two to conduct manual verification on submitted data).

Table 3. Various services' enrolment processes, their design factors and consensus based perceived workload

	ItR[1]	*ItG*[2]	*I*[3]	*D*[4]	*Perceived Workload (by consensus)*
Hotmail	10	2	No	No	28%
Gmail	8	2	No	No	13%
Facebook	6	1	No	No	0%
National e-ID	NA	3	Yes	Yes	81%
PayPal[5]	13	1	Yes[6]	Yes[7]	78%
Skype	11	2	No	No	5%

[1]Items to Recall [2]Items to Generate [3]Interruptions to daily routines [4]Delays [5]Including credit card verification [6]User needs to get hold of a bank statement [7]Can take several days until transaction is visible in a credit card/bank statement

Sensitivity of NASA-TLX

During the follow-up sessions, students were also individually asked to go through a number of fictitious enrolment tasks for fictitious e-services. These tasks are based on common enrolment process configurations generally used in e-government services. A number of e-services from around the world were surveyed and for each service's enrolment-process the researcher recorded its ItR, ItG, I and D values. This afforded the researcher the possibility to construct a set of fictitious tasks, based on real-world services with increasing levels of identity assurance requirements and workload (see Table 4). Table 5 shows how these nine tasks map onto real-world e-services.

Table 4. Set of nine enrolment tasks generalised from a survey of commonly found design configurations across various e-services (from low to high workload and assurance levels)

Design factors	**Fictitious enrolment tasks**								
	Low workload			*Medium Workload*			*High Workload*		
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
ItG	0	1	1	2	2	3	4	3	3
ItR	1	2	5	4	5	6	6	9	NA
D	No	No	Minor[2]	Major[3]	Major[4]	Minor[5]	No	Minor[6]	Major[8]
I	No	No	No[1]	No	Yes[4]	No	No	No	Yes[7]

[1]Credit card details are required [2]Wait a few minutes for activation email [3]Wait three days before account is activated [4]Visit closest outlet to confirm identity [5]Upload a recent photo [6]Call free-phone to activate account [7]Visit enrolment office during specific opening hours [6]Three day waiting period till an activation PIN is received by post

Table 5. Examples of real-world e-services adopting enrolment processes similar to the ones presented in Table 4 (as at 2013)

Task	*Based on…*
A	Directorate of labour (Iceland)
B	Estonian e-government portal (Estonia)
C	Birth certificates (Ontario)
D	Comune di Milano (Italy)
E	Student finance (England)
F	Study permits (Canada)
G	Inland revenue (Italy
H	Access key registration (Canada)
I	e-ID registration (Malta)

For each task, a NASA-TLX evaluation was carried out. It was decided to maintain the final NASA-TLX pairwise rating and thus generate a weighted workload rather than a raw TLX score (see Table 6 for resulting weighting values). It is evident from the weighting exercise that digital natives consider Frustration (F), Physical Demand (PD), Temporal Demand (TD) and Effort (E) as the major sources of workload (in this order). Frustration (F) was presented as a measure of irritation, stress and annoyance during the task while Effort (E) was explained to be the level of mental and physical work required to accomplish the task. This corroborates with the consensus based perceived workload levels shown in Table 3 whereby the highest workload scores were given to those enrolment processes that interrupted the primary task. In the National e-ID case students had to visit an office in Valletta, while in PayPal's case participants had to wait a couple of hours or days until a small PayPal transaction was processed and made visible on the credit/debit card statement. The transaction details on the statement contain an activation code which is required to complete the verification process (i.e., to confirm card ownership).

Table 6. Workload dimension weighting by students following the final pairwise comparison

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	0.7	3	2.7	1.7	2.5	4.4
Median	1	3	3	1	3	5
Std.Dev	0.8	1.4	1.1	1.2	1	1

The participants' overall weighted workload values for each of the nine fictitious enrolment processes presented during this session are shown in Table 7.

Table 7. Median value for the mean weighted workload (MWW) score across all participants for the nine fictitious enrolment processes

Task	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
MWW	0%	0%	18%	11%	32%	14%	12%	21%	81%

A series of tests using the Related-Samples Wilcoxon Signed Rank non-parametric test for non-normally distributed data were carried out to determine whether there is a statistically significant difference between reported workload levels and corresponding tasks designed to be incrementally demanding. The null hypothesis set for these tests is that no statistically significant increase in perceived workload exists between tasks that are designed to be incrementally demanding. In some cases, although the task was intended to be less demanding than the subsequent one, it turned out that digital natives perceived it as more demanding; although a fairly low statistical significance is reported. For example, tasks C and D as well as tasks F and G whereby the null hypothesis was retained.

This can be explained by referring to the participants' supervised workload dimensions' weighting values (see Table 6) wherein Physical Demand (PD) and Temporal Demand (TD) (both given a weight of 3) are considered to be two major contributors to workload, as opposed to Mental Demand (MD) (weight of 1). Although tasks C and F are less demanding than their subsequent tasks (D and G respectively), with lower levels of mental demand (MD), they present users with more physical (PD) and temporal demands (TD) (i.e., travelling, looking up information and waiting for account activation).

Table 8. Tests to determine whether there is a statistically significant difference between reported workload levels for tasks designed to be incrementally demanding

Null Hypothesis[1]	*Significance (.05)*	*Decision*
PEW[*] for Task C over Task B	.001	Reject the NH
PEW for Task H over Task G	.033	Reject the NH
PEW for Task I over Task H	.003	Reject the NH
PEW for Task C over Task A	.001	Reject the NH
PEW for Task F over Task E	.039	Reject the NH
PEW for Task H over Task B	.001	Reject the NH
PEW for Task G over Task C	.039	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets is equal to 0 (i.e., there is no statistically significant increase in perceived workload for subsequent incrementally demanding tasks).

[*] *PEW*: Perceived Enrolment Workload

Supervised vs unsupervised NASA-TLX

Consider Tables 1, 2 and 6. The weighting values for some of the workload dimensions provided via the online questionnaire (unsupervised) are considerably different from those provided for the same dimensions during the follow-up sessions (supervised) – see Table 9 for a synthesis of results. This presents the possibility that participants who had no immediate supervision, as opposed to the supervised group, may have interpreted the rating scales differently from the supervised group. If this is the case, the unmodified (original) NASA-TLX process would not be suitable in an unsupervised environment and with untrained participants. A set of tests are presented below to assess this hypothesis

Table 9. Workload dimension weighting (median) varied when students were supervised as opposed to unsupervised responses (i.e., no immediate help was available)

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Unsupervised (online)	4	0	2	3	2	4
Unsupervised (online without e-ID)	3	1	3	2	2	4
Supervised (follow-up sessions)	1	3	3	1	3	5

Given a non-normal distribution for the workload dimensions' weighting, a set of non-parametric tests were conducted using the Related Samples Wilcoxon Signed Rank test to determine whether there is a statistically significant difference between the unsupervised and supervised sets of weighting values (see Table 10). The following null hypothesis was therefore adopted: the median of differences between each pair of data sets (e.g., Supervised MD and Unsupervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two).

Table 10. Tests to determine whether there is a statistically significant difference between an Unsupervised and a Supervised TLX weighting exercise (i.e., pairwise comparison)

Null Hypothesis[1]	*Significance (.025)*[2]	*Decision*
Supervised MD and Unsupervised MD	.000	Reject the NH
Supervised PD and Unsupervised PD	.000	Reject the NH
Supervised TD and Unsupervised TD	.304	*Retain* the NH
Supervised OP and Unsupervised OP	.021	Reject the NH
Supervised E and Unsupervised E	.011	Reject the NH
Supervised F and Unsupervised F	.000	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets (e.g., Supervised MD and supervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two)

[2] A comparison of two tests under different conditions is being presented using a Bonferroni adjusted alpha level (0.05/2 = 0.025)

Raw TLX vs mean weighted workload

Table 11 shows the medians for MWW and Raw TLX workload (RTLX) together with their respective deviations from the mean. RTLX does not take workload dimensions' weighting into consideration and is calculated by dividing the sum of all workload dimensions' raw ratings for each task/participant by six, the total number of dimensions. Eliminating this final pair-wise comparison to generate the MWW may in turn simplify the TLX process even further. To test this hypothesis a Spearman's rho correlation was run on the non-normally distributed values for MWW and RTLX. Two tests were carried out, one on the data collected during the follow-up workshops (117 observations from 13 participants reporting on nine fictitious tasks) and another test on values reported through the online questionnaire (94 students who had to enrol for an e-ID before using the e-service). In both cases the Spearman's rho revealed a positive and statistically significant relationship between MWW and RTLX (r_s [117] = .989, $p < .001$ and r_s [94] = .937, $p < .001$ respectively). In line with these observations, Cao et al. (2009) observed that RTLX

is more commonly adopted over MWW, citing the high correlation between weighted and unweighted workload scores as the main determining factor.

Table 11. This table shows the set of nine fictitious enrolment tasks together with their respective median MWW values alongside the median RTLX values

Task	*MWW*	*St. Dev.*	*RTLX*	*St. Dev*
A	0%	2.2	0%	2.6
B	0%	9.7	0%	8.1
C	18%	15.6	16%	13.3
D	11%	21.8	8%	17.7
E	32%	28	33%	22.3
F	14%	18.6	13%	17.1
G	12%	9.1	13%	8.8
H	21%	13.3	21%	13
I	81%	27.3	72%	24.4

Discussion

The use of NASA-TLX to measure perceived workload in the exam registration process and e-ID enrolment (where applicable), provided the author with very useful insights. This, together with data from follow-up sessions, helped to understand how students related to NASA-TLX's terminology and processes, as originally introduced by Hart and Staveland in (1988), with the aim to maximise NASA-TLX's validity and useability for this group of users and within this context.

Workload manifests itself in different ways

Students who have used the exam registration e-service, but had to go through the e-ID enrolment process, were expected to give significantly higher overall workload ratings than those who already had an e-ID, mainly due to the additional physical and temporal workload involved in travelling and queuing. However, this was not found to be the case, as there is a negligible difference in overall MWW between the two groups. By drilling down into NASA-TLX's multi-dimensional results it was noticed that sources of workload were significantly different for the two groups. Both presented a high measure of overall workload, albeit for different reasons. In principle those who had to enrol for an e-ID were concerned with delays and interruptions to their primary task; however, they indicated that the exam registration process was – in comparison – acceptable. On the other hand, students who already had an e-ID based their feedback mainly on the non-functional aspects of the exam registration process, such as lack of clarity in the process and site loading speed, resulting in a high level of frustration. Uni-dimensional workload measurement techniques do not explain the user experience in its entirety. Issues in design and performance can cause frustration, and this can be an equally important contributor to perceived workload, together with the more traditionally accepted sources of workload such as the physical and cognitive demands. The author recommends the adoption of a multi-dimensional workload assessment tool in order to understand the various sources of workload for different service alternatives. Future governments depend on the trust of younger citizens, and the interaction with

government institutions is formative for trust perceptions. Riegelsberger and Sasse (2010) point out that trust depends on the users' perception of motivation and competence – so being confronted with less than competently designed e-government services will undermine young people's trust in government.

Demystifying workload dimensions

Although provided with on-screen guidelines, participants in follow-up sessions were at times confused while rating certain dimensions, especially Own Performance (P), Effort (E) and Temporal Demand (TD). In particular Temporal Demand (TD) caused a level of confusion in its interpretation. Participants were often confused if Temporal Demand refers to how long it took to complete the task, or how long it should have taken.

Temporal Demand (TD) was originally introduced in NASA-TLX as a measure of time related pressure during a task, specifically the pace at which tasks occurred. This is a very context specific dimension especially suited for critical scenarios such as an emergency landing of an aircraft in bad weather. As is, this dimension may not be adequate for non-critical and mundane tasks. Further to this, some participants also voiced their concern on the similarity of certain workload dimensions. They explained:

> The main problem is that some of them are really similar. And you wouldn't know what to choose.

It was a non-trivial task to help participants understand the difference between the more abstract workload dimensions such as: Frustration (F) and Own Performance (P) or Effort (E) and Mental Demand (MD). Students were given the opportunity to think aloud and clarify their doubts throughout the exercise by asking questions. As one participant said:

> The only thing which struck me was the 'own performance' rating. Sometimes it is a bit hard to figure out what you did right or wrong so it's kind of hard to assess own performance.

Another comment related to how participants felt while conducting the final pairwise comparison, especially when they were asked to choose between Physical (PD) and Mental Demand (MD):

> Participant A: I also feel lazy with my choices.
>
> Participant B: True, true, same here.

In this case, both participants felt uncomfortable disclosing the fact that they preferred mental demand rather than physical demand; therefore, it can be seen that lack of anonymity may influence feedback. This ties in with Malheiros's (2014) observations on disclosure, whereby participants are less likely to disclose information comfortably and honestly if it portrays them in a bad light.

Keep out of reach of digital natives?

A series of tests, presented in Table 10 indicate that a supervised TLX exercise will yield a significantly different result in the way the six workload dimensions are weighted by digital natives. In the follow-up sessions the facilitator explained each and every workload dimension before going through the different tasks. This might have contributed towards the variance in interpretation, and thus in weighting outcomes, between online and workshop participants. Table 9 shows the differences in the interpretation of rating scales with and without supervision.

It was noticed that this group of users did not fully understand the official NASA-TLX descriptions for the various workload dimensions, in particular those for Mental Demand (MD), Effort (E) and Own Performance (P). Specific and age-appropriate examples were found to be helpful.

NASA-TLX, e-government enrolment and digital natives — does it really work?

Can this technique be used to measure workload confidently with digital natives? This section will tackle a subset of tasks from the nine fictitious enrolment processes presented during the follow-up sessions and their respective workload ratings across the six dimensions. Statistical tests show that there is a significant correlation between the resulting ratings and the demands imposed by the task. Figure 6 represents the overall mean adjusted ratings for three of these fictitious tasks, across the six workload dimensions. Service D had no major workload issues; however Temporal Demand (TD) and Frustration (F) were rated as being considerably high as the task required three days for account activation. Service G had low levels of workload across all dimensions; however, Mental Demand (MD) was the highest rated dimension. This can be explained by the fact that participants had to come up with a new password, a password hint and a call-in PIN to be used to authenticate themselves in case they need to call a help-desk. Service *I* had the highest ratings across all dimensions, and this was especially evident in Physical Demand (PD), Temporal Demand (TD), Effort (E) and Frustration (F). Half a day of travelling and queuing is required to complete the identity verification process as well as a three day period until the activation PIN is received by post.

Table 12. This table shows three different tasks from the set of nine fictitious enrolment tasks – denoting the participants' perceived mean weighted workload (MWW)

Task	*ItR*[1]	*ItG*[2]	*I*[3]	*D*[4]	*MWW*
D	4	2	No	Major[5]	11%
G	6	4	No	No	12%
I	NA	3	Yes[6]	Major[7]	81%

[1]Items to Recall [2]Items to Generate [3]Interruptions to daily routines [4]Delays [5]Wait three days before account is activated [6]Visit enrolment office during specific opening hours [7]Three day waiting period till an activation PIN is received by post

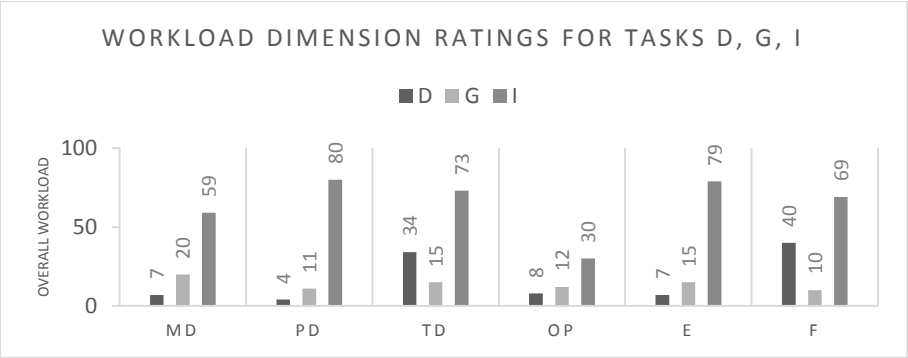


Figure 6. This chart shows the overall mean workload for the three tasks listed in Table 12

A degree of consistency was observed between the perceived workload for enrolment processes used on popular online services (see Table 3) and median weighted workload values for the nine enrolment tasks for fictitious services (see Table 7). Some noticeable examples are provided in Table 13. Although the two sets of results are close, one cannot exclude the possibility of other design factors placing significant influence on workload, especially on dimensions such as Frustration (F) and Effort (E).

Table 13. Contrasting perceived enrolment workload (PEW) derived by consensus from actual enrolment processes with TLX-based Mean Weighted Workload (MWW) values for similar, but fictitious tasks

Real Service	*PEW*	*Fictitious Service*	*MWW*
Hotmail	28%	Task H	21%
Gmail	13%	Task G	12%
National e-ID	81%	Task I	81%

Furthermore, following a series of tests presented in Tables 10 and 11, it was determined that even though the nine fictitious tasks were presented in a random order, on average participants reported statistically significant differences in perceived workload for tasks designed to be more demanding.

A final set of tests sheds more light on the need to retain the pairwise comparison exercise that is used to produce weighted workload values for each participant. Results provided in Table 11 show that there aren't any major advantages for the adoption of MWW values over RTLX, given the additional effort required from participants to complete the final pairwise comparison. Eliminating this final step may in turn simplify the TLX process even further.

Modifying NASA-TLX for use in e-service enrolment

The meaning of Temporal Demand (TD) may need to be modified to fit within an e-government context. The experience of 'feeling rushed' may not be an appropriate measure for enrolment processes, as opposed to other situations such as engaging

landing gears during an emergency landing. In the follow-up sessions Temporal Demand (TD) was expressed as a measure of the time required to complete the task. The associated hint should read: 'How much time did you require to complete this task?' This represents the perceived amount of time taken-up by the enrolment portion of an e-service, rather than the pressure exerted from time limitations.

Simpler definitions and context specific examples are needed for most of the rating scales:

- *Own Performance*: How confident were you during the enrolment process? Was the process easy to follow?
  The inverted labels for Own Performance (Good to Poor rather than Low to High) did not seem to be problematic.
- *Physical Demand*: How much physical effort did the process involve? Did you have to search for some documents? Did you need to go somewhere in-person to complete the transaction?
- *Mental Demand*: How much thought was required during this process? Did you have to come up with new secrets, such as usernames, passwords or PINs? Did you have to provide a lot of information to complete the form(s)?
- *Effort*: Considering both mental and physical demand, did it require a lot of effort to perform the process?
- *Frustration*: How irritating or annoying was this enrolment process?

If possible provide a channel for immediate feedback during the TLX rating process using voice over IP (VoIP) if physical proximity is not possible. Finally, Raw TLX was found to be a suitable measure to inform designers about the perceived workload for this group of users (digital natives), while also simplifying the overall rating process. This was mainly due to the fact that a high level of correlation was found between Raw TLX and MWW values, making the additional effort required to generate MWW values unjustifiable.

Conclusions

Following a rigorous empirical exercise, this paper offers insights on the applicability of NASA-TLX as a highly-cited human factors technique to measure the impact of enrolment process design on e-government service users. The literature reviewed positions NASA-TLX as one of the better workload assessment techniques, in both sensitiveness and ease of use. It has been adopted in a number of domains and applications, from analysing flight crew complement requirements and down to evaluating software interfaces. This study's goal was to shed more light on the effectiveness of NASA-TLX, particulary when used by and on digital natives in an e-government context.

NASA-TLX provided interesting insights into the possible sources of workload for this group of users, and it was found to be fairly sensitive to changes in workload parameters, informing the researcher of possible actions to reduce workload perceptions, improve adoption and if compulsion exists, minimise resentment. With minor modifications NASA-TLX could be improved to serve its purpose better

within this particular context and with this user group. This also includes additional guidance on the meaning and implications of the various workload dimensions. Finally, it has been noted that in this context there are no major advantages arising from the use of the MWW metric over RTLX.

References

Axelsson, K. & Melin, U. (2012). Citizens' attitudes towards electronic identification in a public e-service context an essential perspective in the eid development process. In Hans J. Scholl, Marijn Janssen, Maria A. Wimmer, Carl Erik Moe, and Leif Skiftenes Flak, editors, *Electronic Government*, volume 7443 of Lecture Notes in Computer Science, pp.260–272. Springer Berlin Heidelberg.

Cain, B. (2007). *A review of the mental workload literature*. Technical report, Toronto, Canada: Defence Research and Development.

Cao, A., Chintamani, K.K., Pandya, A.K., & Ellis, R.D. (2009). NASA-TLX: Software for assessing subjective mental workload. *Behavior Research Methods*, 41(1):113–117.

Card, S.K., Moran, T.P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Commun.ACM*, 23(7), 396–410.

GARTEUR. (2003). Action Group FM AG13. *Garteur handbook of mental workload measurement*. Technical report.

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

Hart S.G. & Staveland L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology - Human Mental Workload*, *52*,139– 183.

Hill S.G., Iavecchia H.P., Byers J.C., Bittner A.C., Zaklad, A.L., & Christ, R E. (1992). Comparison of four subjective workload rating scales. *Human Factors, 34*, 429-439.

Inglesant, P. & Sasse, M.A. (2007). Usability is the best policy: Public policy and the lived experience of transport systems in London. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not as We Know It* - Volume 1, BCS-HCI '07, pp. 35–44. British Computer Society.

Katsanos, C., Karousos, N., Tselios, N., Xenos, M., & Avouris, N. (2013). KLM form analyzer: Automated evaluation of web form filling tasks using human performance models. In *Human-Computer Interaction - INTERACT 2013*, volume 8118 of Lecture Notes in Computer Science, pp. 530–537. Springer Berlin Heidelberg.

Malheiros, M. (2014). User behaviour in personal data disclosure. PhD thesis, University College London.

Porter, C., Sasse, M. A., & Letier, E. (2012). Designing acceptable user registration processes for e-services. In Proceedings of *HCI 2012 - The 26th BCS Conference on Human Computer Interaction*. BCS.

Porter, C., Sasse, M. A., & Letier, E. (2013). Giving a voice to personas in the design of e-government identity processes. In Proceedings of *HCI 2013 - Research*

to Design: Challenges of Qualitative Data Representation and Interpretation in HCI. BCS.

Prensky, M. (2001). Digital natives, digital immigrants. *On the horizon*, *9*(5), 1–6.

Riegelsberger, J. & Sasse, M. A. (2010). Ignore these at your peril: Ten principles for trust design. In *Trust 2010. 3rd International Conference on Trust and Trustworthy Computing*.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*, 61–86.

Sasse, M. A., Steves, M., Krol, K., & Chisnell, D. (2014). The great authentication fatigue - and how to overcome it. In P.L.Patrick Rau, editor, *Cross-Cultural Design*, volume 8528 of Lecture Notes in Computer Science, pp 228–239. Springer International Publishing.

Steves, M., Chisnell, D., Sasse, M. A., Krol, K., Theofanos, M., & Wald H. (2014). Report: Authentication diary study. NISTIR 7983. *Technical Report NISTIR 7983*.

“What does beep mean?” – context free interpretation of short sinus wave stimuli

Matthias Wille, Sabine Theis, Peter Rasche, Christina Bröhl, Rebecca Kummer, & Alexander Mertens
Institute of Industrial Engineering and Ergonomics, RWTH Aachen University
Germany

Abstract

Although human-machine-interaction had become more technically mature through the last decades many devices still communicate with simple sinus beeps. Those beeps can be differentiated based on rhythm and tone pitch, but they are very abstract signals, which can be only interpreted by knowing the device very well or if the manual was studied profoundly. Therefore, it would be interesting if there is a kind of common structure in the beep appearance that is interpreted context free by all humans. For instance, fast sequences of beeps indicate an alarm or ascending pitch means a question while descending pitch means a statement (like typical human intonation). Hereby, possible age effects – based on technical generations and according experiences – should be taken into account. In our study 13 younger and 13 elderly participants were confronted with 8 sinus beeps that differ in rhythm and pitch and had to rate each of them on 5 continuous scales: alarm - note, question - confirmation, pleasant - unpleasant, important – unimportant, distinct – inconclusive. Results showed significant effects on 3 of the 5 scales meaning that a common interpretation in these dimensions exists. This can be important for designing auditory signals based on sinus waves.

Introduction

During the last decades technology and therefore also human-machine interaction has developed rapidly. Displays have become smaller, brighter and with higher resolution. Speech interaction has left its niche and is now possible on most smartphones. And with the internet of things machines have improved to communicate with themselves: A refrigerator can now order milk by itself without contacting its human master if it registers a lack of milk. But although communication with machines has developed so far, most of the machines still communicate with simple sinus tone sequences with their human masters for information input or input confirmation. These beeps can be sent by a washing machine, car radio, medical devices, smartphone, computer or anything else. The reason for these beeping machines may be the simple and unexpensive feasibility of a sinus tone generator which at least do the job to catch the human attention. The human operator might know the meaning of this beep-signal, because he is used to

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

the machine or eventually (but unrealistic) he has red the manual. However, in most cases we only know that the machine wants something from us, but not the concrete content. This study investigates a common interpretation in beep sequences depending on pitch variation and speed.

The study took place within the "Tech4Age" Project (www.tech4age.de). Within this project we develop a pattern language of how human machine interaction should be designed for elderly users in healthcare context. In context of another study on multimodal perception (Wille et al., 2016) the question emerged, if sinus tone sequences alone can have some common meaning or interpretation that influence the perception and action of participants.

Theoretical background

Searching the literature about interpretation of sinus tone sequences no study or description can be found with that *concrete* topic. For sure, already Helmholtz (1863) had focused on sound perception and the research on auditory displays in common has a long history: Adam and Tucks (1976) for instance tested reaction time to different warning signals in different ambient noise enviroments. They found out that reaction time to these signals do not only refer to their intensity but also to their structure: Reaction to a "wail" sound (a slow variation of frequency between 400 and 925 Hz over several seconds) was much slower than to a "Yeow" sound (descending change in frequency from 800 to 100 Hz every second). But the faster reaction to more intense sounds is only true for simple reaction tasks and not for choice reaction tasks as Van der Molen and Kuess (1979) found out. That illustrates, that the structure of a sound has influence, but have to be seen in interdependency with the task. In common it is known that audicons which simulate meaningfull sound are better for HCI than abstract earcons (Brewster et al., 1993; Gaver, 1986; Blattner et al., 1989 and many more). Garzonis et al. (2009) compared the effectiveness of earcons and audicons in term of their intuitiveness, learnability, memorability and user preference and found out that audicons significantly perform better than earcons across all four measures. If using abstract earcons Edworthy et al. (2011) have found that the learnability of a set of earcons is greatly enhanced by avoiding similar temporal patterns and increasing the range of type of sounds.

Many other aspects of sound has been investigated, even the pitch of sounds as a perceptual analogue to odor quality (Belkin et al., 1997). But if it comes to simple sinus tones and the interpretation of their sequence in a common, inherent and context free way literature is missing or at least was not found.

However, sinus tone sequences have been investigated about their emotional implication (Scherer & Oshinsky, 1977). Descending melodies are associated with pleasantness and upscending melodies are also interpreted as being pleasant as long as the pitch variation is big enough. This stands in line with Smith and Cuddy (1986) who found out, that sinus tone sequences are interpreted in general as more pleasant as more melodic they are. So it can be seen as proved that sinus tone sequences have a common emotional interpretation. Therefore the question pops up if there is also a common rational interpretation of sinus tone sequences, like a sequence is interpreted as more or less urgent signal.

Method

To investigate if simple sinus tone sequences are interpreted in the same way from different people a laboratory-study was conducted where 8 sinus tone sequences with up to three tones in up to three pitches were randomly presented on a smartphone and participants rated their impression on 5 scales immediately after hearing the tone.

Participants

26 subjects participated in the study with an age of 16-76 years. The sample was divided by median split into two age-groups to investigate possible age effects. In the younger group were 13 subjects between 18-28 years (Mean = 22.38, SD = 2.434, 7 male /6 female). In the older group were 13 subjects aged 49-76 years (Mean = 64.23, SD = 9.391, 4 male / 9 female). All participants had normal or corrected to normal sight and normal hearing abilities.

Material – the sinus tone sequences

Each stimulus consisted of 5 time-units which were 100 msec long and were filled with a tone or a pause where no sound is played. So overall each stimuli was ½ second long. The tones were pure sinus waves with three different pitches that were matched harmoniously (C6 - 1046.50 Hz; G5 - 783.99 Hz; C5 - 523.25 Hz). Furthermore, all pitches were selected in an area where no influence of presbycusis is to be expected. Figure 1 shows the structure and the naming of the stimuli: Each stimulus was assigned by a 5-digit number which reflects the sequence of tones and pauses. A pause was indicated by 0, the lowest pitch (C5) by 1, the middle pitch (G5) by 2 and the highest pitch (C6) by 3. Longer tones in a sequence (like in 11033 or 22222) are played continiuosly without any break or new attack envelope in between.

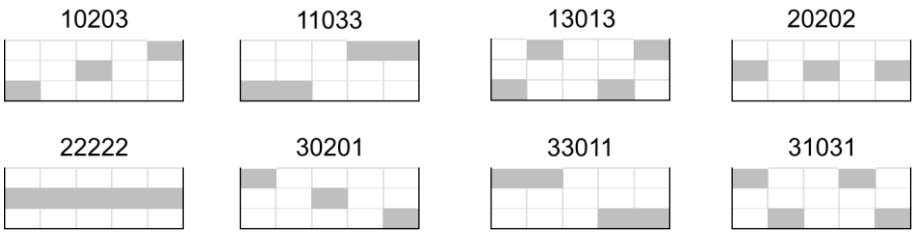


Figure 1. Structure and naming of the sinus tone sequences.

Scales

Participants had to rate each stimulus on 5 continiuos scales. These scales were based on expert interviews and a short evaluation regarding intelligibility with participants during pretesting (these participants are not part of the sample described here). The scales are listed below with the original German dimensions in brackets and followed by a brief explanation which was given to the participants also during introduction.

- alarm - note („Alarm – Hinweis“)
  An alarm calls for fast action while a note is not time critical.
- question - confirmation („Frage – Bestätigung“)
  A question requires a timely response, while during a confirmation no response is required (for example, confirmation-tone after a command entry)
- pleasant – unpleasant („Angenehm – Unangenehm“)
  A signal can be perceived as pleasant or unpleasant by you.
- important – unimportant („Wichtig – Unwichtig“)
  A signal can be perceived as important or unimportant by you.
- distinct – inconclusive („Eindeutig – Uneindeutig“)
  If the stimulus could be clearly assigned to the presented scales it is distinct if differentiation is more difficult, then the stimulus is inconclusive.

Material –the experimental application and used hardware

The experimental setup was built as a native Android app and presented on a Nexus 5 smartphone running Android 6.0. The volume was fully turned up to ensure all participants perceived the stimuli clear and with the same volume. Figure 2 shows a screenshot of the application. Sounds were played once by pressing the “next” button on the lower right, but can be repeated as often the participant want by pressing the “repeat” button on the top right. Scales were presented as continuous horizontal faders with both endpoints named in german language. By default the middleposition was displayed for each new stimulus. Participants rated the scales by moving the fader into the desired position and each scale which was already rated became grey (like the first two in figure 2). Not yet rated scales were shown in black (as the last three in figure 2). Participants had to rate on all scales before the application allowed going to the next stimulus by pressing “next”. If not all scales were rated a message box popped up when pressing “next” and told to rate all scales of the actual stimulus first. This was to ensure that all participants rate all stimuli and not skip some scales by accident or by disinterest. If participants prefered the middle position given in the beginning they can rate it that way by just touching the scale for a short time and bringing the signifier back into the middle position. The internal resolution on each scale was 0-1 with 3 decimal places which was transferred to 0 – 1000 afterwards.

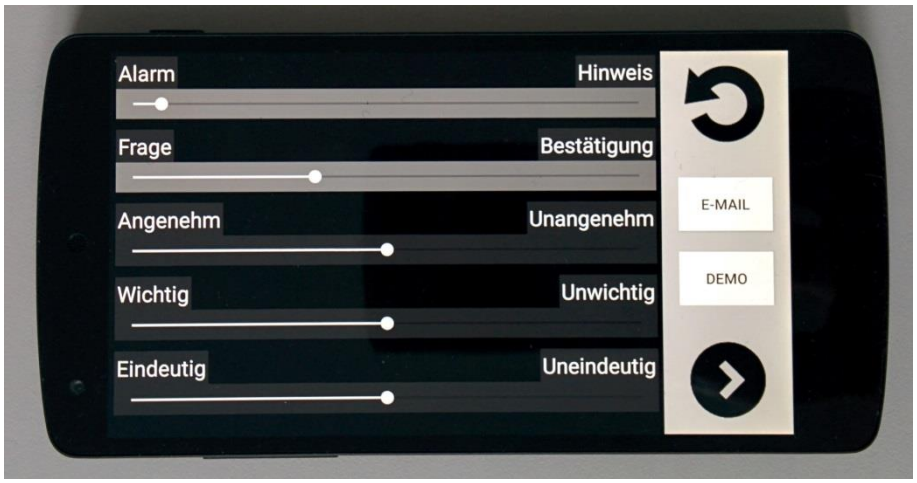


Figure 2. Screenshot of the experiment app.

Procedure

This study was conducted as an appendix to another study about multimodal perception (Wille et al., 2016; more to come), where participants had to categorize multimodal stimuli consisting of visual, audio and tactile material wether they were rhythmic or not. The sinus wave stimuli in this previous experiment matched the middle high of this study (G5 – 783.99 Hz) but no variation in pitch was given and the focus layed on all three dimensions, not only on tones. Between the last stimuli of the previous study and the first trials of this study a break of about 15 minutes was given. As the previous study focused on fast reaction to any stimuli and contained no pitch variation or rating of stimuli an influence on this follow up study might be controllable.

When starting the here described study, participants first got an instruction of what they will have to do followed by an introduction to the scales, where a written description of each scale was given to them alike the description given here a subchapter above. Finally, stimuli were played once in random order to familiarize the participants with the set of stimuli. The experiment started after participants had the chance to pose remaining questions. During the study the investigator monitored the hand position of the subjects and corrected them if they cover accidently the speakers of the presentation phone. Conducting the experiment was accounted for 15 minutes.

Experimental design and variables

This study followed a repeated measurement plan, where each participant rated each stimulus once on all scales. Independent variable was the stimulus – 8 different sinus tone sequences. The age of subjects was a between subject factor (median split into 2 groups). Main dependend variables were the ratings of each stimulus on the 5 scales. As participants had the choice to repeat the stimulus as often as they want during their rating, the number of repetitions is interpreted as a factor of uncertainty:

If one stimulus is repeated more often than others participants had more difficulties to rate it. The rating time was measured from the beginning of stimulus presentation until the rating was finished and is interpreted as a factor of difficulty. All data was analysed in SPSS 22 using Anovas with repeated measures.

Results

The result section initially contemplates the number of stimulus repetitions and the time for rating the participants needed. Subsequently, on each scale differences between stimuli and age-group were reported. This is followed by a correlation matrix to see if the scales intercorrelate. Finally an overview is given for the ratings of each stimulus on the significant scales.

Repetition count and rating time

A repeated measurement Anova with stimulus as independent variable, age-group as between subject factor and repetition count as dependent variable showed no effect for stimulus [$F(4.69, 112.62) = 0.773$, $p = .564$, $\eta p^2 = .031$], age-group [$F(1, 24) = 0.848$, $p = .402$, $\eta p^2 = .029$] or the interdependency of those facors [$F(4.69, 112.62) = 0.641$, $p = .659$, $\eta p^2 = .026$]. Mauchly's test indicated that the assumption of sphericity had been violated for the stimuli [$X^2(27)$ 56.03, $p < .001$], therefore Greenhouse- Geisser corrected tests are reported ($E = .67$). Overall each stimulus was repeated less than one time (0.71) with a maximum of 6 repetitions for one stimulus.

As a factor of difficulty the rating time was measured from the beginning of stimulus presentation until the rating was finished. A repeated measurement Anova with stimulus as independent variable, age-group as between subject factor and rating time as dependent variable showed no effect of stimulus [$F(3.99, 38.38) = 0.811$, $p = .522$, $\eta p^2 = .037$] but an significant effect of age-group [$F(1, 21) = 8.204$, $p = .009$, $\eta p^2 = .281$]: older participants took about 36 seconds to rate while younger participants required about 22 seconds. An interdependency between age and stimulus was not found [$F(3.99, 38.38) = 0.972$, $p = .972$, $\eta p^2 = .006$]. Again Mauchly's test indicated that the assumption of sphericity had been violated [$X^2(27)$ 71.65, $p < .001$] therefore Greenhouse- Geisser corrected tests are reported ($E = .67$).

Scales

The ratings on scale 1 alarm – note showed an significant effect for stimulus [$F(7, 168) = 2.313$, $p = .028$, $\eta p^2 = .088$] but no effect of age-group [$F(1, 24) = 1.086$, $p = .308$, $\eta p^2 = .043$] or interdenpendency between stimulus and age-goup [$F(7, 168) = 0.974$, $p = .452$, $\eta p^2 = .039$]. Here Mauchly´s test indicated no violation of sphericity. The scale alarm – note is shown in figure 3.

Scale 2 question – confirmation showed no significant effect: neither for stimulus [$F(4.32, 103.62) = 1.602$, $p = .175$, $\eta p^2 = .063$], nor for age-group [$F(1, 24) = 1.751$, $p = .198$, $\eta p^2 = .068$] or interdependency of both factors [$F(4.32, 103.62) = 1.515$, $p = .199$, $\eta p^2 = .059$]. Here Mauchly's test indicated that the assumption of sphericity

had been violated [$X^2(27)$ 49.80, p = .005] therefore Greenhouse- Geisser corrected tests are reported (E = .62).

Scale 3 pleasant – unpleasant showed a significant effect for stimulus [F(7, 168) = 5.710, p < .001, ηp^2 = .192], but not for age-group [F(1, 24) = 0.634, p = .434, ηp^2 = .026] or interdependency [F(7, 168) = 0.720, p = .655, ηp^2 = .029] (see figure 4).

Scale 4 important – unimportant showed again a significant effect for stimulus [F(4.65, 111.61) = 3.529, p = .006, ηp^2 = .128], but no effect for age-group [F(1, 24) = 0.094, p = .761, ηp^2 = .004] or interdependency of stimulus and age [F(4.65, 111.61) = 1.684, p = .149, ηp^2 = .066] (see figure 5). On this scale Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(27)$ 50.70, p = .004, therefore Greenhouse- Geisser corrected tests are reported (E = .66).

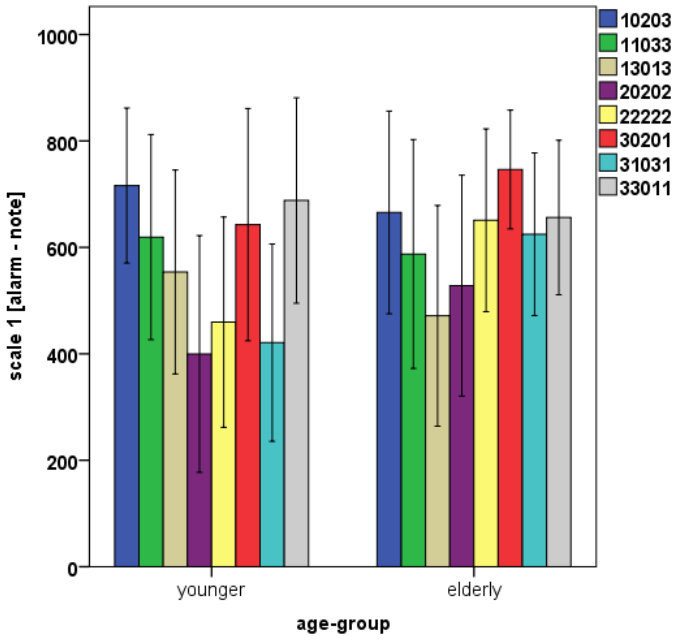


Figure 3. Rating of 8 stimuli at scale 1 alarm – note. Error bars reflect the 95% confidence interval.

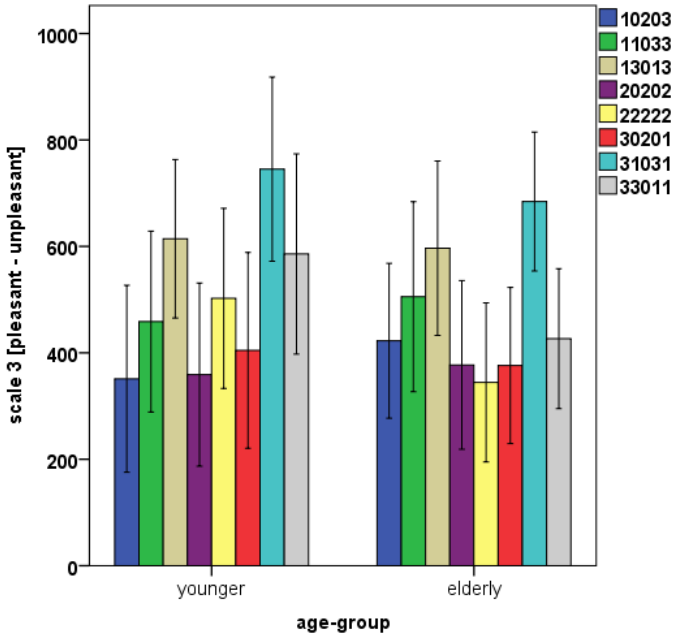


Figure 4. Rating of 8 stimuli at scale 3 pleasant – unpleasant. Error bars reflect the 95% confidence interval.

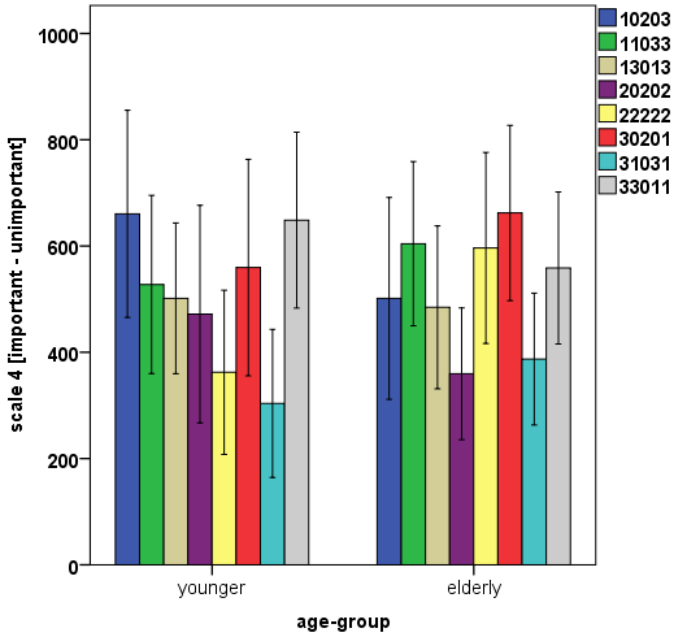


Figure 5. Rating of 8 stimuli at scale4 important - unimportant. Error bars reflect the 95% confidence interval.

The scale 5 distinct – inconclusive showed no significant effect at all: Not for stimulus [$F(7, 168) = 0.820$, $p = .572$, $\eta p^2 = .033$], nor for age-group [$F(1, 24) = 0.008$, $p = .930$, $\eta p^2 = .000$] or interdependency of both factors [$F(7, 168) = 1.095$, $p = .369$, $\eta p^2 = .044$].

To sum it up: Two scales failed to show significant differences between the stimuli: scale 2 (question - confirmation) and 5 (distinct - inconclusive). In that scales no pattern could be found that indicate a common interpretation. But three scales (1 alarm – note; 3 pleasant – unpleasant; 4 important – unimportant) showed significant effect of stimulus, which means that participants have a common way to rate the stimuli in that dimensions.

Correlations between scales

Table 1 shows the correlations between the scales. These correlations are based on the ratings of all participants across all stimuli (N = 26 participants * 8 stimuli = 208). As table 1 shows the scales do highly intercorrelate, meaning the ratings are not independent from each other. Focusing on the scales that showed a significant effect (1,3,4) it can be said that stimuli that are rated rather as note than alarm (high score on scale 1) are associated with being pleasant (low score on scale 3, with negative correlation) and less important (high score on scale 4). On the other hand alarms are more unpleasant and important. Although these associations make sense in real life it means statistically that the alpha risk of the Anovas for each scale is enlarged (because asking several times for the same phenomen) and has to be corrected from 5% to 1% (divided by the number of scales or times asking for the same phenomen). In that case scale 1 would be no more significant, while scale 3 and 4 still hold their significance.

*Table 1. Correlations among scales (** p < .001)*

	Scale 1	*Scale 2*	*Scale 3*	*Scale 4*	*Scale 5*
Scale 1	-	.184**	-.220**	476**	.026
Scale 2	.184**	-	-.073	.189**	-.075
Scale 3	-.220**	-.073	-	-.099	.107
Scale 4	.476**	.189**	-.099	-	.193**
Scale 5	.026	-.075	.107	.193**	-

Interpretation of stimuli

Figure 6 shows the mean ratings on the two remaining significant scales for each stimulus. Age-group as factor was dropped as no age effects showed up during the analysis of each scale. The stimuli “10203”, “30201” and “20202” are interpreted as being most pleasant, while “13013” and “31031” are being perceived as most unpleasant. That means that pauses between the tones – independent from pitch variation – are more pleasant to the subjects, while fast pitch jumping without any pauses in between is perveived as unpleasant. The most important stimuli were “20202” – the one which was also identified mostly as alarm – and “31031” – the one which was the most unpleasant. The most unimportant stimuli were “10203”,

"11033", "30201" and "33011" – all of the stimuli with a slow pitch variation separated by a pause.

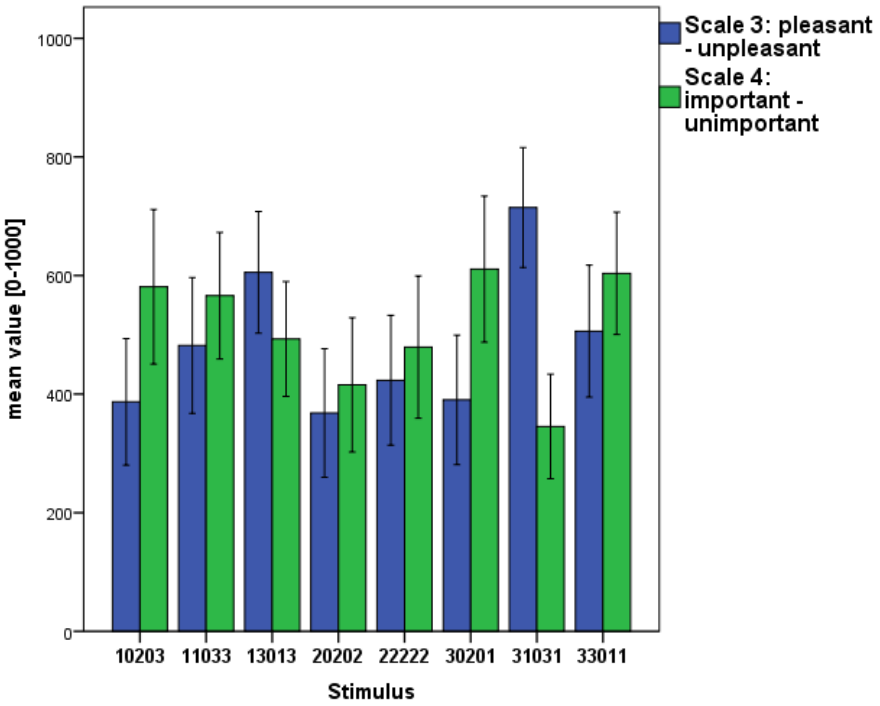


Figure 6. Rating of 8 stimuli at the two significant scales: Scale 3 (pleasant – unpleasant) and 4 (important – unimportant). Error bars reflect the 95% confidence interval.

Discussion

This study followed an explorative setting to investigate if simple and abstract sinus wave sequences do carry an inherent and context free meaning if applied in human computer interaction. The results have to be interpreted as relative within the set of given stimuli. This set was very minimalistic regarding number of stimuli and also regarding variation within those stimuli to identify certain factors that influence the perception of simple sinus wave sequences.

Only two of five scales showed significant effects: Scale 3 (pleasant – unpleasant) and 4 (important – unimportant) and the effect size is rather small. However, even with that given restricted set of only 8 stimuli and 26 participants it is now proved that the sinus wave stimuli are interpreted in the same way from participants in these dimensions. This effect is independent from age or cohort as young and elderly participants showed the same reaction. This does not only stand for a "design for all" approach but also once again it can be seen that the signals are interpreted consistently the same way by both groups. So it can be nomore denied that even abstract sinus wave signals contain context free characteristics that will influence the interpretation and reaction to these signal. But it is too early to build common rules

for that interpretation based on pitch and sequence of the stimulus. Therefore more studies have to be done with a wider range of stimuli.

The present finding that pitch variation is interpreted to be more pleasant than stimuli without pitch variation stands in line with the findings of Scherer and Oshinsky (1977) and Smith and Cuddy (1986). Additional we can say that pitch variation separated by pauses is interpreted as more unimportant signal compared to signals without pitch variation or without pauses. However, scales are highly intercorrelated, which means they are not independent and seem to charge on a concept we do not know yet and which is somehow a mix of unpleasant importance on one side of the scale and pleasant unimportance on the other side of the scale. This has to be further evaluated in future research, where scales have to be validated and ideally be renamed in one global concept or other independent scales that do not correlate have to be found. But although the behind concept is not fully clear yet, this work has proven that there is some global, context free interpretation of abstract sinus wave signals, that will influence the perception of and reaction to these sounds. And as long as machines use those sounds (which are not ideal at all, see theoretical background) understanding this concept is crucial for developing auditory displays.

Acknowledgement

This publication is part of the research project "TECH4AGE", which is funded by the German Federal Ministry of Education and Research (BMBF, Grant No. 16SV7111) supervised by the VDI/VDE Innovation + Technik GmbH.

References

Adams, S.K., & Trucks, L.B. (1976). A procedure for evaluating auditory warning signals. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 20, No. 8* (pp. 166-172). Sage Publications.

Belkin, K., Martin, R., Kemp, S.E., & Gilbert, A.N. (1997). Auditory pitch as a perceptual analogue to odor quality. *Psychological Science*, *8*, 340-342.

Blattner, M.M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human–Computer Interaction*, *4*, 11-44.

Brewster, S.A., Wright, P.C., & Edwards, A.D. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 222-227). ACM.

Edworthy, J., Hellier, E., Titchener, K., Naweed, A., & Roels, R. (2011). Heterogeneity in auditory alarm sets makes them easier to learn. *International Journal of Industrial Ergonomics*, *41*, 136-146.

Garzonis, S., Jones, S., Jay, T., & O'Neill, E. (2009). Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1513-1522). ACM.

Gaver, W.W. (1986). Auditory icons: Using sound in computer interfaces. *Human-computer interaction*, *2*, 167-177.

Helmholtz, H. (1863). Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik: Mit in den Text eingedruckten Holzschnitten. J. Vieweg.

Molen, M.V.D., & Keuss, P.J.G. (1979). The relationship between reaction time and intensity in discrete auditory tasks. *The Quarterly journal of experimental psychology*, *31*, 95-102.

Scherer, K.R., & Oshinsky, J.S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, *1*, 331-346.

Smith, K.C., & Cuddy, L.L. (1986). The pleasingness of melodic sequences: Contrasting effects of repetition and rule-familiarity. *Psychology of Music, 14*, 17-32.

Wille, M., Seinsch, T., Kummer, R., Rasche, P., Theis, S., Bröhl, C., Mertens, A. & Schlick, C. (2016). Development of an Experimental Setup to Investigate Multimodal Information Representation and Superposition for Elderly Users in Healthcare Context. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 104-115). Berlin: Springer International Publishing.

Spatially distributed visual, auditory and multimodal warning signals – a comparison

André Dettmann & Angelika C. Bullinger,
Technische Universität Chemnitz,
Germany

Abstract

Spatially distributed warning signals are able to increase the effectiveness of Advanced Driver Assistance Systems. They provide a better performance regarding attention shifts towards critical objects, and thus, lower a driver´s reaction time and increase traffic safety. The question which modality is used best, however, remains open. We present three driving simulator studies (30 participants each) with spatially distributed warnings, whereby two focused on spatial-visual as well as auditory warnings respectively. The third study, which combined the most promising approaches from the previous studies, depicts a multimodal spatial warning system. All studies included a baseline without secondary tasks and warnings. Afterwards, subjects were confronted with multiple (30+) critical objects while performing a secondary task. The chronological order of warnings was randomly mixed between spatial, non-spatial and no warning during the first two studies. Data from reaction times, eye tracking data, and questionnaires were collected. Results show that spatial-visual directed warnings are more effective than non-spatial warnings in large distances, but subjects do have difficulties in detecting objects in peripheral regions when they are distracted. While auditory spatial warnings are not as efficient as literature implies, it still performed best in this particular situation. Results of the multimodal warning study, discussion and implications on Advanced Driver Assistance Systems (ADAS) conclude the paper.

Introduction

In 2015 2,516,831 people were involved in traffic accidents (German Federal Statistical Office, 2015). In relation to that 305,659 people were injured and 3,459 died. That was an increase of +1.1% to the prior year. Human errors while driving were the main reason for accidents (68.8%). They were either distracted, ignored traffic regulations, made mistakes at turning manoeuvres, or failed to keep an appropriate safety distance to vehicles driving ahead. This behaviour is the consequence of human errors in perception and cognition of information and responding appropriately in such situations (Bubb, 1993; Spanner-Ulmer, 2008). Lee (2008) concluded that accidents happen when the driver fails to look at the right time at the right object. Posner (1980) stated that by means of spatial cues a faster reaction towards spatial stimuli is possible.

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Using such spatially directed cues in Advanced Driver Assistance Systems (ADAS) can help to support drivers in these domains (Jentsch, 2012) by directly focusing (or shifting) the attention towards hazardous objects. In that way the risk potential for distracted drivers can be minimised. Possible warning signs could be visual or auditory cues. Spatially directed visual cues are able to shift the driver's attention towards relevant objects using a novel Head-Up-Display (HUD) based on light emitting diodes (LED) mounted underneath the windscreen. This can result in lower reaction times and benefits the overall quality of reaction in critical situations (Dettmann et al., 2014). Due to technical limitations there might be restrictions when presenting optical warning signals with LED-HUD in the peripheral field of vision (e.g. driver is looking to the right and the spatial-visual cue is on the left outside of the HUD). To help to counteract the problematic of warning perception in the peripheral visual field a second warning modality, auditory cues are promising. Similar to the visual warning there are also two presentation modes possible: a conventional, undirected auditory warning signal and a spatially directed auditory warning signal. They seem to be a beneficial integration as audible cues are generally more efficient in shifting attention compared to visual cues (Proctor et al., 2005; Scott & Gray 2008; Haas & van Erp 2014). Using spatially directed sound is based on the assumption that selective spatial attention can improve cognition (Lampar, 2011; Haas & van Erp 2014).

To examine the concept and the effectiveness of spatially directed cues for each visual and auditory warning modality, we undertook two driving simulator studies. The first using visual warnings (spatially directed and undirected) by means of the LED-HUD and a second study using solely auditory warnings in the same manner. The main focus was to investigate the impact of the warning concepts on reaction times. Auditory and visual cues were then compared and glance behaviour was analysed depending on the warning modality (auditory vs. visual). The following third study embraced the most promising approaches from the previous studies towards a combined multimodal warning. This approach was taken, as it was the intention to examine the distinct effects and limitations of each spatially distributed visual and auditory warning modality on drivers' perception. The third study primarily evaluates the overall concept of a spatial warning system using a novel LED-HUD.

The present paper initially describes the realisation of visual and auditory directed cues, including the underlying warning concept, followed by the description of the simulator study and the representation of results for each study. This also includes the results of a sector based analysis of where potentially dangerous objects appeared (near/far, inside/outside, left/right). Discussion and implications on Advanced Driver Assistance Systems with spatial warning signals will conclude the paper.

Realisation of spatially directed warning cues

For the realisation of the visual warning an LED-HUD connected to the driving simulator was used. The used LED-HUD can be regarded as a multifunctional human-machine-interface for advanced driver assistance systems or in-vehicle information systems applicable for collision warnings (Lindner et al., 2009),

attentional control (Kienast et al., 2008) or as an assistant system for autonomous driving. The LED-Panel is built as a 135° circle segment mounted underneath the windscreen. It is able to cover 50° of the driver's field of view. It consists of nine panels each with 256 LEDs and is able to display spatially directed optical warnings. These can either be semantically enriched pictograms (e.g. direction, distance or object types) or simple flashes. Both signal types can be presented on the LED-HUD towards the position of a potentially dangerous object. For the present studies a white flashing square was used as shown in figure 1. This warning concept was evaluated in a prior study by Pöschel et al. (2013). Ten experts compared symbols representing different types of objects (e.g. car or pedestrian) or information (e.g. traffic signs).

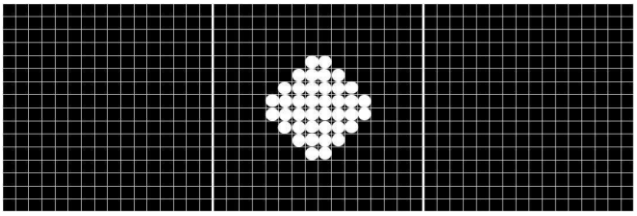


Figure 1. Warning concept on the LED-HUD

Despite the low semantic information, the symbol was favoured because of its high contrast and the possibility to implement high blinking frequencies (e.g. flashes) which are particularly useful in peripheral regions as they are highly sensitive for flicker (Mühlstedt, 2013). The used flash was visible for 250 ms long in a 0.4 s cycle.

The auditory warning used in the second study was developed based on the guidelines of design principles for auditory warnings and were tested in the driving simulator (Wogalter & Leonard, 1999; DIN EN ISO 15006, 2011). Four experts evaluated issues concerning volume and speaker position. The four point audio system in the simulator was positioned in a way that hearing and detecting the direction of audible cues was possible which was also tested and approved by the expert group. The audio signal had a base frequency of 2,573 Hz and was 60 dB loud and contained four repeated single tones.

A multimodal approach was examined for the third study. The same LED-HUD and warning was used (spatially directed visual warning) as well as the presented auditory signal (conventional warning design, undirected).

Method

To examine the effectiveness of spatially directed cues, two conditions were presented to the subjects: first, a spatially directed cue oriented towards an object (occurring stationary vehicle; see figure 2, right), whereas the second condition was the conventional, undirected cue where the alert occurs in the middle (figure 2, left). A "no cue" condition was equally presented in the first study to examine the general effectiveness of warning signals. Each participant underwent both conditions in balanced order while being distracted by a secondary task. The third study with the

multimodal warning design contained only one warning condition. To distract the participants from driving, i.e. looking at the street, a secondary task was used, consisting of moving geometric figures. The underlying concept is to recreate a situation where the driver is not aware of his situation and possible hazardous objects. Therefore, the overall effectiveness of each warning modality can be evaluated free of any confounding variables. The subjects were asked to recognise size, length or colour features and report their answer to the experimenter. The secondary task was meant to be a contrary cue with no relationship to the primary driving task (Schweigert, 2012). Before the actual experiment, all subjects completed a familiarisation and a baseline drive. The baseline drive excluded the secondary task to measure the reaction times when the driver was not distracted.

All studies were conducted on a simulated straight test track measuring 5000m by 150m without any traffic. The goal was to minimise any possible distractions caused by the environment or traffic.

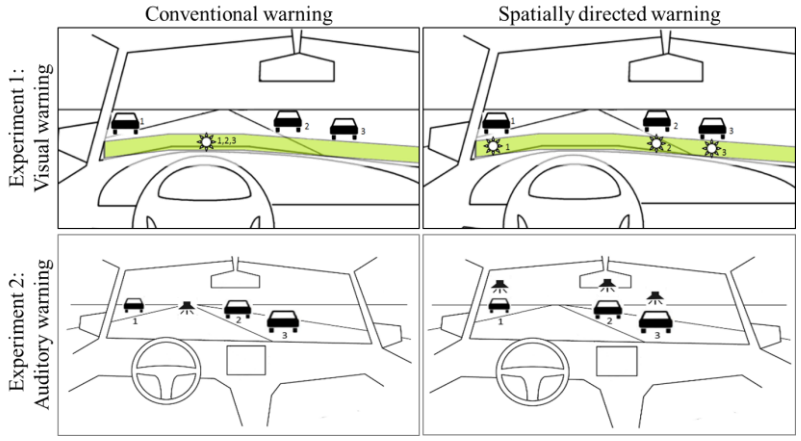


Figure 2. The warning designs as used in the studies

While driving, 32 static objects were presented to the subjects. They appeared in random intervals of 5 – 10 s at a recommended speed of 50 km/h. The distance between the driver and the object was randomised from 40m to 200m. Furthermore, the point of occurrence varied between eight sections. These sections consisted of four horizontal sections divided into close and long-range sectors. Distances up to 100m from the subject represented the close range, while distances from 120m upwards formed the far range. No objects were presented at distances from 100m to 120m to achieve a higher separation effect. Objects along the lane were mirrored and up scaled to the right side due to a wider opening angle of the windscreen on the right. Objects became smaller and less noticeable at larger distances.

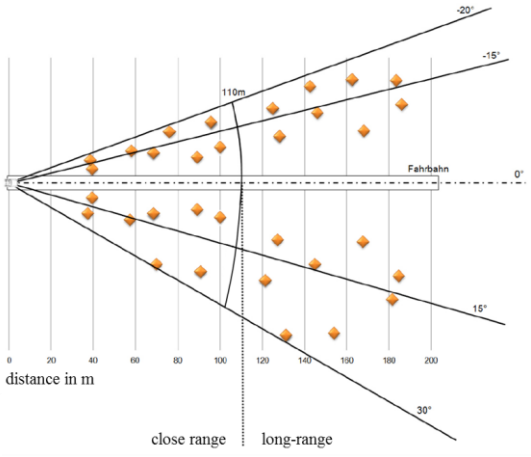


Figure 3. Distribution and position of potentially critical objects in eight sectors

To measure the quality of the warning signals (e.g. reaction time), participants were requested to press a button on the steering wheel as soon as they detected the object (either the left button when the object occurred on the left side or the right button when it occurred on the right). Correct responses were followed by some visual and audible feedback as the object disappeared when the correct button had been pressed. Also, a distinct signal tone different from the warning signal gave feedback. The experiment was conducted as a within-subjects design. To measure the reaction times driving data was gathered (simulator software SILAB 4.0). This data includes time markers for object appearance and button actuation. The reaction time is defined as the difference between the occurrence of the objects and the moment when the participants pressed the button on their steering wheel. Therefore, the difference between those markers was calculated. Additionally, eye-tracking data (system Dikablis) was gathered to control in which direction the subjects were looking in the moment of the objects appearance. This was done by comparing the time stamps of the driving and the eye-tracking data and then looking at the glance direction of the subjects.

To ensure comparable reaction times the data points had to meet the following criteria: A data point is valid if (1) participants looked at the secondary task at the time the object occurred, (2) the object was perceived and the correct steering wheel button was pressed and (3) the buttons on the steering wheel worked perfectly. All studies combined a total of 2157 valid data points collected from 90 participants who were distracted by a secondary task. This results in an average of 719 valid data points (76.7 %) for each study. The most common reason for non-valid data points was the participants' attention which was not centred at the secondary task (19.7 %). The value 1.9 % resulted from pressing the button on the wrong side of the steering wheel. 1.7 % was caused by technical problems with the buttons on the steering wheel. Non-valid data points were excluded from the calculations and, therefore, had no influence on the results. Table 1 gives an overview of the valid data points for each study.

Table 1. Overview of valid samples

STUDY	*Valid samples [N]*	*Not processes secondary task [%]*	*Problems with buttons [%]*	*Wrong reaction [%]*
Visual	695	20.2	5.7	0.2
Auditory	722	19.8	0.0	1.2
Multimodal	740	19.1	0.0	3.9
Overall average	721.3	19.7	1.9	1.7

For all valid data points a sector based and a field based analysis (near/far, inside/outside, left/right) was conducted to get better insights of how effective each warning modality in certain areas performs. Figure 4 gives an overview of where the sectors are located.

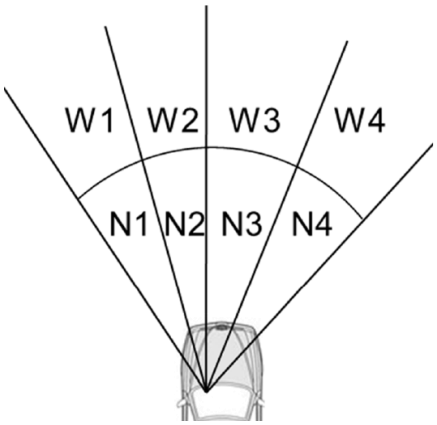


Figure 4. Location of each zone for the sector based analysis

Additionally, in all studies the participants were to answer non-validated questionnaires helping to evaluate their current mental condition (five-point Likert items: excited/nervous, awake/tired), system acceptance (intention to buy), sociodemographic data (age, gender, visual and hearing performance) as well as the design of the cues (five-point Likert items: e.g. urgency, locatability, aggressiveness).

Sample

Participants of all three simulator studies were matched for age and gender being comparable to the sample of Dettmann et al. (2014). Each study counted 30 participants aged from 19 to 45 (M_{age} = 28.7, SD_{age} = 6.3), whereby 36.7 % were female. All participants owned a valid driving licence and drove an average of 11,063 km per year. The mean age of the sample was equivalent to Dettmann et al. (2014). Significant differences were only found in annual mileage (analysis of variance (ANOVA), $F(2, 55) = 3.972$, $p = .022$) whereas the "auditory" and the

"multimodal" group hold significantly fewer driven kilometres per year than the "visual" group.

Table 2. Overview of the participants

STUDY	*#*	M_{age}	SD_{age}	$M_{annual\ km}$
Visual	30	30.4	7.6	18,133
Auditory	30	28.9	4.8	8,661
Multimodal	30	26.9	6.0	9,021

Results

For all studies the influence of the mental condition on valid data points can be neglected. There was no relationship found between the drivers' condition (excited/nervous, chi-squared test: $\chi^2(3, N = 2820) = 4.87, p = .18$) and the percentage of valid data points. The same applies for the condition "awake/tired" ($\chi^2(4, N = 2820) = 2.96, p = .56$). For all ANOVA conducted, the Levine's Test for Homogeneity of Variance met the assumption of equality of variances.

Comparison of visual warning conditions and no warning

For the sector based analysis of the reaction times for the visual warnings a single factor variance analyses was carried out. To examine the overall effectiveness of the LED-HUD all visual warnings (directed and undirected) were compared with the "no warning" condition. This resulted in significant lower reaction times for the warning condition in the sectors W1-W4 (see table 3). For close objects (sectors N1-N4) only for sector N2 a significant difference was found between all directed and undirected visual warnings and the "no warning" condition.

- W1 ANOVA, F(2, 83) = 32.938, p < .001
- W2 ANOVA, F(2, 89) = 60.976, p < .001
- W3 ANOVA, F(2, 82) = 23.469, p < .001
- W4 ANOVA, F(2, 97) = 32.469, p < .001
- N2 ANOVA, F(2, 61) = 6.467, p = .003

In sector W2 a significant difference between the directed and undirected warning was found (post hoc analysis (Scheffé's method) $p = .034$) while in sector W4 only a mild trend was found (post hoc analysis (Scheffé's method) $p = .088$). This applies to the sector N3 as well (ANOVA, $F(2, 65) = 4.509, p = .015$). Through the post hoc analysis also a trend between directed and undirected warning was found (Scheffé's method p = .093). The reaction times for all three conditions are presented in table 3 and figure 5.

Table 3. Overview of the reaction times for the visual warnings

Visual warning		W1	W2	W3	W4	N1	N2	N3	N4	All
Spatially directed	*M*	1.41	1.11	1.23	1.25	1.08	1.12	0.91	1.00	1.14
	SD	0.44	0.37	0.41	0.30	0.36	0.54	0.22	0.27	0.45
Spatially undirected	*M*	1.24	1.64	1.59	1.66	1.05	0.92	1.22	1.03	1.26
	SD	0.53	0.75	0.70	0.68	0.36	0.38	0.46	0.27	0.57
Baseline	*M*	2.47	3.60	3.36	2.87	1.10	1.56	1.33	1.06	0.76
	SD	0.78	1.21	1.84	1.08	0.46	0.76	0.66	0.35	0.20

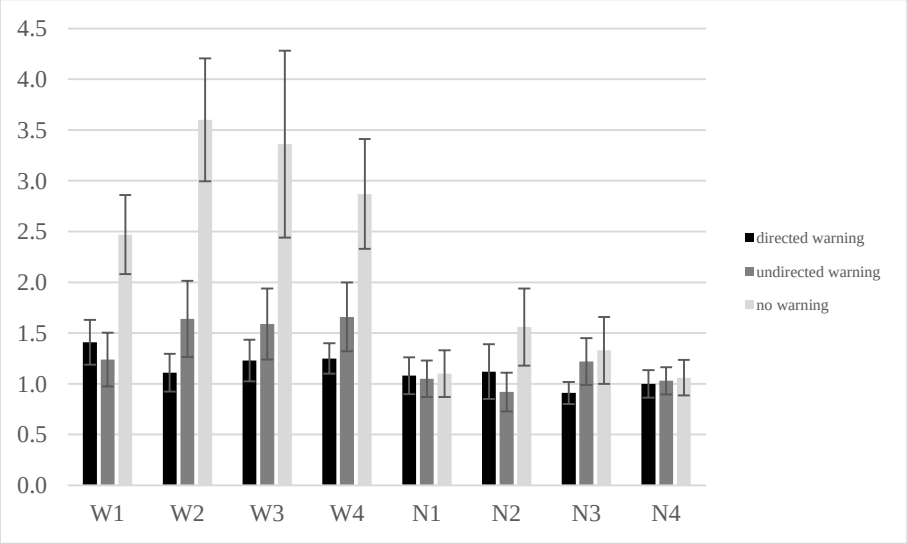


Figure 5. Reaction times for the visual warning

When combining the sectors into fields (near/far) and (left/right) only for near and far fields significant differences were found:

- near ANOVA, $F(2, 325) = 6.890$, $p = .001$
- far ANOVA, $F(2, 360) = 116.834$, $p < .001$

The post hoc analysis showed no differences for the near sectors, but for the far sectors: directed (M = 1.25 s; SD = 0.39 s) and undirected (M = 1.59 s; SD = 0.7 s) warning (p = .024). Regarding the left and right fields also significant differences were found:

- left ANOVA, $F(2, 333) = 58.287$, $p < .001$
- right ANOVA, $F(2, 352) = 46.940$, $p < .001$

While there is no difference between directed and undirected warnings to the left, on the right side a significant difference between directed (M = 1.07 s; SD = 0.32 s) and undirected (M = 1.44 s; SD = 0.64 s) warnings (p = .014) was found.

From the questionnaires it was found, that 21 out of 30 subjects rated the directed visual warnings as very supportive. Only seven rated the undirected warnings more useful and two subjects found that they felt no support from the warnings. Figure 5 is showing three additional subjective ratings to characterise directed or undirected warnings on a five-point Likert scale.

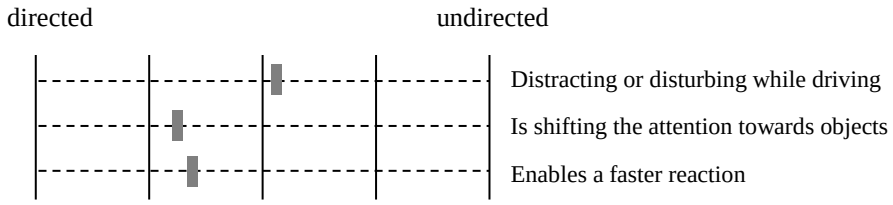


Figure 6. Subjective rating of the visual warnings

Comparison of auditory warning conditions

For the second study the same sector and field based analysis (see figure 4) was carried out. Overall no significant differences between directed and undirected warnings were found ($t(720) = 1.71$, $p = .088$). Significant differences between directed and undirected warnings were found in the sectors W3 ((t-test) $p < .001$), N4 ($p < .001$) and N1 ($p < .001$). For the sectors W3 and N4 undirected auditory warnings performed better and for sector N1 directed warnings showed faster reaction times. All reaction times regarding auditory warnings are shown in table 4 and figure 7.

Table 4. Overview of the reaction times for the auditory warnings

Auditory warning		W1	W2	W3	W4	N1	N2	N3	N4	All
Spatially directed	*M*	1.01	1.19	1.48	1.40	0.91	0.9	0.93	1.15	1.11
	SD	0.34	0.62	0.49	0.71	0.23	0.32	0.43	0.32	0.49
Spatially undirected	*M*	0.99	1.11	1.10	1.22	1.33	0.92	0.93	0.84	1.06
	SD	0.34	0.43	0.36	0.57	0.24	0.35	0.43	0.26	0.41

When analysing the fields (near/far), (inside/outside) and (left/right) significant differences for the far sectors were found (ANOVA, F(3, 198) = 8.371, p = .000). The undirected warning signals ($M = 1.10$ s, $SD = 0.44$ s) were more efficient than directed signals ($M = 1.26$ s, $SD = 0.59$ s). For the combined left and right sectors inconsistent results were found. On the left directed auditory warnings worked better (ANOVA, F(1,364) = 3.964, p = .047; directed: $M = 1.01$ s, $SD = 0.43$ s; undirected: $M = 1.09$ s, $SD = 0.38$ s) and for the right side undirected auditory warning signals had significant lower reaction times: ANOVA, F(1,356) = 15.633, p < .01; directed: $M = 1.22$ s, $SD = 0.53$ s; undirected: $M = 1.02$ s, $SD = 0.44$ s. No difference were found for the (inside/outside) fields. Figure 8 is showing three subjective ratings to characterise directed or undirected auditory warnings on a five-point Likert scale.

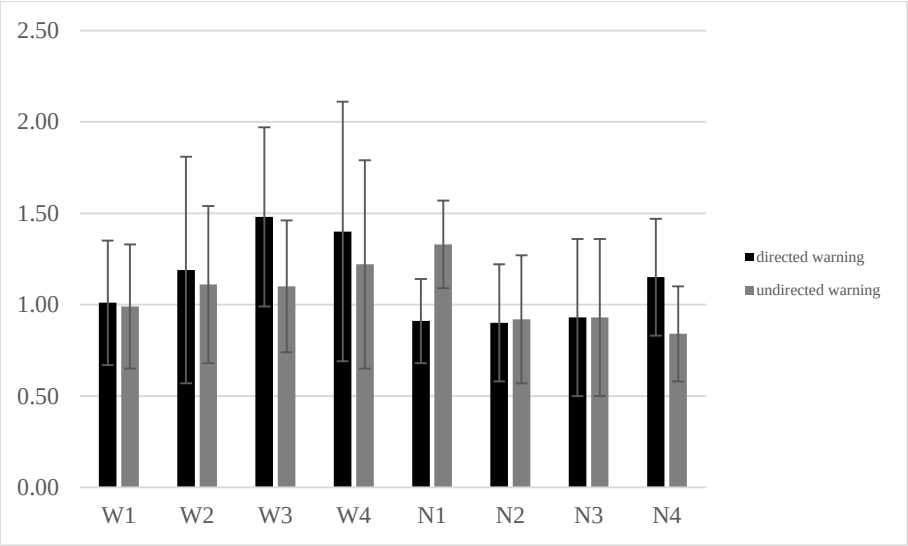


Figure 7. Reaction times for the auditory warnings

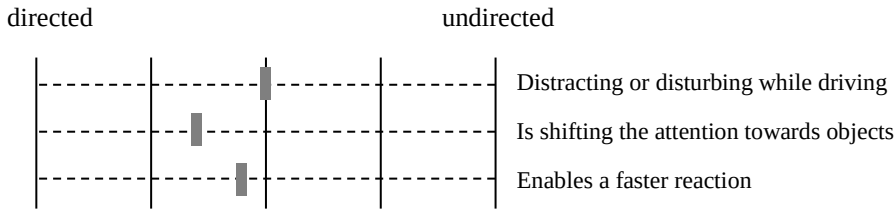


Figure 8. Subjective ratings for the auditory warnings

Results of the multimodal study

In the third study only one condition, the multimodal approach, was examined. The warning design embraced the most promising approaches (visual directed and auditory undirected) from the previous studies towards a combined multimodal warning. The single factor variance analyses reports no differences between each sector ($F(2, 7) = 24.8$, $p < .001$). The comparison between all three conditions is showing that the multimodal warnings are performing significantly better across all sectors. Table 5 and figure 9 are giving an overview of the multimodal reaction times compared to the visual directed and auditory undirected warnings and their respective reaction times.

Table 5. Overview of the all reaction times

STUDY		W1	W2	W3	W4	N1	N2	N3	N4	All
Spatially directed visual	*M*	1.41	1.11	1.23	1.25	1.08	1.12	0.91	1.00	1.09
	SD	0.44	0.37	0.41	0.3	0.36	0.54	0.22	0.27	0.45
Conventional auditory	*M*	0.99	1.11	1.10	1.22	1.33	0.92	0.93	0.84	1.06
	SD	0.34	0.43	0.36	0.57	0.24	0.35	0.43	0.26	0.37
multimodal	*M*	0.76	0.78	0.78	0.77	0.76	0.76	0.74	0.71	0.76
	SD	0.31	0.2	0.18	0.16	0.22	0.29	0.14	0.17	0.20

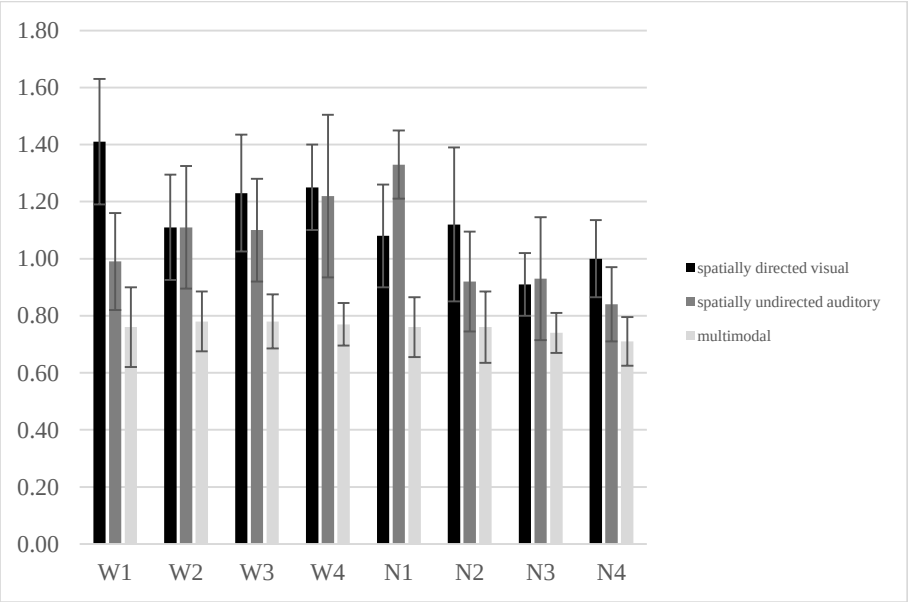


Figure 9. Reaction times for all three conditions

Discussion

In accordance with other research works (Fricke, 2009; Scott & Gray, 2008) the present experiment compared reaction times with and without warning. In general the efficiency of warnings can be confirmed repeatedly. Visual warnings represented on a LED-HUD as used in the first experiment are suitable for shifting back the attention to objects on the roads when the driver is distracted. Especially spatially directed visual warnings are helping to shift the attention towards relevant objects in road transport quickly and intuitively. Best results were found for object detection and reaction in far-distant sectors (W2 – W4). If the spatial directed warning is displayed far left (e.g. in sector W1) it becomes disadvantageous as the driver may not recognise the warning signals. The processing of the secondary task and the resulting orientation to the right an increased effectiveness of the directed warning compared to the undirected warning could not be proved. This is also true for all near-field areas. Regarding the questionnaires, it also becomes apparent that the directed warning signals were rated more useful than undirected warnings. Subjects

find, that the directed warning is more supportive in shifting the attention and provides a faster reaction.

For the second study, it was assumed that spatially auditory warnings are able to increase the effectiveness in peripheral regions as acoustical spatial stimuli are omnidirectional and no fixation is needed. It was found that this is true for far-distant sectors (W1) but not for near peripheral sectors (N1). Overall no consistent results regarding reaction times were found that spatially directed auditory warnings are better. When looking into glance transitions, it was found that subjects perform a visual search for the object. The auditory spatial cue is not applied to the object position and hence, shows no potential to shift the attention towards objects compared to undirected auditory warnings. This result seems contrary to what literature implies (e.g. high spatial resolution capability (in Goldstein, 2007); faster reaction towards spatial stimuli (Ho et al., 2006; Lampar, 2011)). Those statements are the results of studies, when the subjects are "active" listeners. In the present studies, hearing and processing auditory stimuli was part of the (very simple) driving task, the processing of the visual secondary task, object perception with the help of a warning and finally a response selection and execution (compare Wickens & Hollands, 2000). Hearing is a much more passive activity during the present studies and therefore, the mentioned benefits seem to lose their effectiveness.

The third study combined best results of both studies (visual directed and auditory undirected) depicts a multimodal spatial warning system. Multimodal warnings with visual directed warnings are highly effective. Reaction times benefit from this warning design across all sectors. In comparison with recent literature the relative reaction time reduction for multimodal warnings compared to uni-modal warnings was similar. Biondi et al. (2017) found even lower reaction times but only for brake reactions and not a visual search task as the present studies demanded. The higher effectiveness by using two redundant warning modalities was also proven by Lees at al. (2012). One limitation mentioned was that in a multimodal setup a visual cue is able to affect the advantages of solely auditory warnings (see Fernandez-Duque & Posner, 1996). For the present studies the visual cue benefits towards the searching task. When looking at the gaze data, it was found, that the auditory signal is shifting the attention back to the road and the directed visual warnings are shifting the attention towards the objects. This has a great meaning for the driving task because the warning design supports the driver at choosing and interpreting important information in his traffic and environmental situations. The chance of misinterpretation of traffic situations could be reduced significantly, especially in case of time constrains, high complexity or missing of relevant information.

One limiting factor for the present studies is, that the test persons are distracted by a secondary task. They were cognitively distracted and constantly aware of a required reaction which typically does not happen under actual conditions. Since the attraction is already known and expected, the accuracy of identification is reduced. It is assumed that an appropriate reaction might take longer under real conditions.

Conclusion

In conclusion it can be said that the demonstrated LED-HUD in its form has a high effectiveness regarding the reduction of reaction times. Directed warnings are beneficial with respect to large distances but disadvantageous regarding warning positions out of sight. While spatially directed auditory signals seem not to be fully capable to counteract this, it was found that they at least provide a better attention shift back to the road than on a particular object positioned in space. The multimodel study proved that the combination of warning design helped to shorten reaction times. This warning design counterbalances the advantages and disadvantages of both the visual and auditory warning designs.

Typical situations where spatially directed warnings can save time and hence, make driving saver and more comfortable are as following: complex situations (within the city or at crossing points), restricted environmental conditions (reduced visibility, darkness) or if the driver is distracted and has to interpret current automotive and traffic situations very fast. An important future application can be autonomous driving when the driver is getting a take-over request from the vehicle. In such situations, the driver needs to be fully aware of the situation around him. A multimodal directed warning implemented in an ADAS supports the driver to regain situational awareness as spatially directed warnings have the potential to enable a faster orientation which is needed to recognize relevant objects and to react on them properly.

For further research advanced technical approaches are possible. An optimised presentation of a directed warning can be realised as an animation of the optical cue could guide the attention toward relevant objects. Another opportunity could be the adaption of the LED-HUD warning to the position of the head. But this requires further technical installations, e.g. a Head-Tracker. Further research is also required regarding the adaption of warning designs and timings to the drivers' (emotional) condition.

References

Bubb, H. (1993). Systemergonomie. In H. Schmidtke, I. Jastrzebska-Fraczek: *Ergonomie - Daten zur Systemgestaltung und Begriffsbestimmungen*. Munich: Germany: Carl Hanser,

Dettmann, A., Jentsch, M., Thieme, C., Lindner, P., Bullinger, A.C., & Wanielik, G. (2014). Wirksamkeit räumlich gerichteter Warnungen unter Anwendung eines LED Head-Up-Displays. In *VDI/VW-Gemeinschaftstagung Fahrerassistenz und integrierte Sicherheit (VDI-Bericht 2223) proceedings (pp. 9 – 20)*. Düsseldorf, Germany: VDI Verlag GmbH.

Biondi, F., Strayer, D.L., Rossi, R, Gastaldi, M., Mulatti, C., Advanced driver assistance systems: Using multimodal redundant warnings to enhance road safety. *Applied Ergonomics*, *58*, 238-244

EN ISO 15006:2011. (2011). Road vehicles - Ergonomic aspects of transport information and control systems - Specifications for in-vehicle auditory presentation (ISO 15006:2011), (pp. 1 - 20). Berlin, Germany: Beuth Verlag

Fernandez-Duque, D., Posner, M.I. (1997) Relating the mechanisms of orienting and alerting. *Neuropsychologia, 35*, 477–486

Fricke, N. (2009). Gestaltung zeit- und sicherheitskritischer Warnungen im Fahrzeug. PhD thesis. Technische Universität Berlin, Germany: Fakultät Verkehrs- und Maschinensysteme.

German Federal Statistical Office (2016). Verkehr – *Verkehrsunfälle, Technical series 8 – series 7*. Wiesbaden: German Federal Statistical Office.

Haas, E.C. & Van Erp, J.B.F. (2014). Multimodal warnings to enhance risk communication and safety. *Safety Science, 61*, 29-35

Jentsch, M., & Bullinger, A.C. (2012). Beurteilung von Eingriffen einer Aktiven Gefahrenbremsung in Real- und Simulatorversuchen. In *Forschungsseminar Innovation & Wertschöpfung proceedings*. Chemnitz, Germany.

Kienast, H., Lindner, P., Weigel, H., Henning, M., Krems, J.F., Wanielik, G. & Spanner-Ulmer, B. (2008). Aufmerksamkeitssteuerung mit räumlich gerichteten Anzeigen bei Fahrerassistenzsystemen. In *5. VW/VDI-Gemeinschaftstagung "Fahrerassistenzsysteme und Integrierte Sicherheit" (VDI Bericht 2048) proceedings*. (pp. 413-424). Düsseldorf, Germany: VDI Verlag GmbH.

Lampar, A.L. (2011). Einfluss zeitlicher Aufmerksamkeits-Orientierung auf die Verarbeitung akustischer Reize. PhD thesis. Düsseldorf, Germany: Mathematisch-Naturwissenschaftlichen Fakultät.

Lee, J.D. (2008). Fifty years of driving research. *Human Factors, 50,*. 521–528.

Lindner, P., Weigel, H., Fardi, B., Henning, M., Kienast, H., Wanielik, G., Krems, J. F. & Spanner-Ulmer, B. (2009). Räumlich gerichtete Anzeigen im Fahrzeug: Realisierung einer Mensch-Maschine-Schnittstelle für ein Fußgängererkennungssystem. In *1. Automobiltechnisches Kolloquium proceedings*. Düsseldorf, Germany: VDI Verlag GmbH.

Mühlstedt, J. Roßner, P. Bullinger, A.C. (2013). Die dunkle Seite des Lichts. Diskomfort durch Flicker bei (LED-)Lichtern im Straßenverkehr in Bezug zu peripheren Flimmerverschmelzungsfrequenzen. In E. Brandenburg, L. Doria, A. Gross, T. Günzler, H. Smieszek (Eds.), *Grundlagen und Anwendungen der Mensch-Maschine-Interaktion*, 10. Berliner Werkstatt Mensch-Maschine-Systeme. (pp. 408-414). Berlin, Germany: Universitätsverlag der TU Berlin.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychologie, 32*, 3 - 25.

Pöschel, K., Fritzsch, T., Richardson, N., Jentsch, M., Lindner, P., Bullinger, A.C., Wanielik, G., & Krems, J.F. (2013). Expertenevaluation einer räumlich gerichteten Kollisionswarnung in einem Head-Up-Display. In Chancen durch Arbeits-, Produkt- und Systemgestaltung - Zukunftsfähigkeit für Produktions- und Dienstleistungsunternehmen – 59. GfA-Kongress *proceedings.* (pp. 445 – 448) Dortmund, Germany: GfA-Press

Proctor, R.W., Tan, H.Z., Vu, K.-P.L., Gray, R., & Charles, S. (2005). Implications of Compatibility and Cuing Effects for Multimodal Interfaces. Mahwah, NJ, USA: Lawrence Erlbaum Associates

Schweigert, M. (2002). Fahrerblickverhalten und Nebenaufgaben. PhD thesis. Technische Universität München, Germany: Fakultät für Maschinenwesen

Scott, J., Gray, R. (2008). A Comparison of Tactile, Visual, and Auditory Warnings for Rear-End Collision Prevention in Simulated Driving. *Human Factors*, *50*, 264 - 275.

Spanner-Ulmer, B. (2008). Mensch-Maschine-Kommunikation. In *46. Deutscher Verkehrsgerichtstag - Arbeitskreis VII: "Belastbarkeit des Fahrzeugführers" proceedings*. Goslar: Germany: Deutsche Akademie für Verkehrswissenschaften e.V. – Deutsches Verkehrswissenschaftliches Seminar

Wogalter, M.S., Leonard, S.D. (1999). Attention Capture and Maintenance. In M. S. Wogalter, *Handbook of Warnings* (pp. 123-148). Mahwah, New Jersey, USA: Lawrence Erlbaum Associates

Performance using low-cost gaze-control for simulated flight tasks

Ulrika Ohlander[1], Oscar Linger[2], Veronica Hägg[2], Linn Nilsson[2], Åsa Holmqvist[2], Sandra Durefors[2], Jens Alfredson[1,2], & Erik Prytz[2]
[1]Saab AB, [2]Linköping University
Sweden

In the current study, interaction using gaze control was compared to computer mouse using the MATB-II (Multi-Attribute Task Battery) environment. The study had two aims; the first was to explore the utility of low-cost technologies in a rapid prototyping and testing environment for aviation. The second aim was to use such an environment to compare a novel interaction device (a low-cost gaze control device) to a familiar interaction device (computer mouse). *Method*: Thirty participants performed two scenarios with each interaction device. The software MATB-II provided simulated flight tasks and recorded performance. Mental workload was assessed by the NASA Task Load Index (TLX) questionnaire after each scenario. *Results***:** The results showed that gaze control resulted in significantly higher overall mental workload than computer mouse. Performance was better with mouse in two of the four tasks. *Conclusions***:** Concerning the first aim, the study demonstrated the value of low-cost technology for initial user testing before using more expensive high-fidelity environments. Concerning the second aim, the computer mouse resulted in better performance and lower mental workload. This may either be due to higher user familiarity with computer mouse interaction or to limitations of the gaze control equipment and insufficient adjustments of the interface design to optimize for gaze control.

Introduction

Fighter pilots in air combat are exposed to both high G-loads as well as high workload. Pilots often need both hands to control the aircraft during fast-paced scenarios. Moreover, during flight with high G-loads it is difficult for the pilot to move their arms and hands to interact with the cockpit interface. Today, the main solution to these situations is HOTAS control, Hands on Throttle and Stick, which give the pilot access to control devices on the joystick and throttle as well as the possibility to interact with tactical systems. To a certain extent, voice control is also available as a solution today. As the flow of information from the system to the pilot is constantly increasing, and the tactical systems gets more complex, the need for the pilot to interact with the system is also increasing. This difference in bandwidth between the computer and the human creates a demand to utilise all possible means of interaction between the pilot and the system, and to investigate and explore new possible enabling technical solutions (Jacob, 1991). The technology for interacting

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

with graphical user interfaces using eye gaze is rapidly improving and eye tracking has now become a feasible interaction method to explore further. Using eye movements to control a pointing device is typically seen as a simple and intuitive task that can be performed with high speed (Drewes, 2010; Penkar, Lutteroth, & Weber, 2013). In a simulator study, O´Connel, Castor, Pousette & Krantz (2012) demonstrated that target designation using eye tracking was beneficial. However, additional research is needed to further explore the benefits and limitations of eye gaze as an interaction method in flight decks. The current study was conducted to investigate the impact of eye gaze as input device on workload and performance in a simulated flight scenario in MATB-II (Multi-Attribute Task Battery).

Simulations of complex environments such as flight decks are often complicated and costly to perform. MATB-II is intended to provide the same type of workload as for a pilot in a real airplane using only a desktop PC and gaming hardware for controls, which means a significantly lower cost. The purpose of this study was to examine gaze-control using a commercially available eye tracker, the Eye Tribe, in this type of low-cost simulated environment. The objectives were to investigate the interaction and measure workload when using eye-gaze in a simulated flight deck, and at the same time explore the possibilities to achieve a low-cost testing environment with MATB-II and commercial gaming controls.

In this study, eye-tracking performance was compared to ordinary computer mouse control. Using a mouse is obviously not how a pilot interacts with cockpit systems, but the participants in the study can be assumed to have a high familiarity with mouse control. Thus, the comparison can be regarded as being between the new method, eye-gaze, and a commonly used controlling device for the target group (Rupp, Oppold, & McConnell, 2013).

Mental Workload and Performance

Gopher and Donchin (1986) define mental workload as a function of the mental capacity that is necessary for information and decision processing, and the capacity available in the current situation. A task that is difficult for one person can be considered easy for another. One explanation of this is experience, which reduces the mental workload (Wickens, Hollands, Banburry, & Parasuraman, 2012). Mental workload is often measured by using subjective methods (Charlton, 1996). Subjective measurements have several practical advantages, such as easy implementation and providing detailed data for perceived load (Eggemeier & Wilson, 1991). The NASA Task Load Index (Hart & Staveland, 1988) is one of the most commonly used subjective methods for measuring workload. Mental workload can affect human performance, such that workload that exceeds the operator's coping capacity can lead to sharp performance degradation (Hancock & Warm, 1989). However, performance can be protected against the increase in workload through, for example, increased effort up to a certain limit (Hancock & Warm, 1989; Hockey, 1997). Svensson, Angelborg-Thanderz, Sjoberg, and Olsson (1997) showed that task difficulty is positively correlated to workload, and that workload negatively correlated with fighter pilot performance in a high-fidelity simulated flight deck.

Method

A repeated measures design was used with the independent variables being the control method and task difficulty, and the dependent variables being workload and performance.

Participants

Thirty participants (19 male, 11 female) were recruited to participate in this study. They ranged in age from 19 to 43 years with a mean age of 24.2 (SD = 5.8). All reported normal or corrected-to-normal vision with no color vision deficiencies. All reported that they normally used their right hand for computer mouse interaction. None had prior experience with MATB-II.

Apparatus

MATB-II (NASA, 2014) is a computer program that simulates tasks in cockpits in order to evaluate operator performance and workload (Comstock Jr & Arnegard, 1992). The original MATB has been modified to MATB-II and the major changes include adaption to later versions of Windows operating systems and new graphical interfaces (Santiago-Espada, Myer, Latorella, & Comstock Jr, 2011). There are four different tasks for the user: System Monitoring Task, Tracking Task, Communications Task and Resource Management Task. The program is mainly developed to simulate tasks that correspond to a pilot's mental workload during flight, but can be used for other purposes such as measuring general multitasking capacity (Aricò et al., 2014; Rupp et al., 2013).

Modifications in the MATB-II interface were made to allow tasks to be performed using gaze control. As the eyes constantly make small flickering movements (saccades) there can be problems using eye control to target small buttons in graphical user interfaces with high precision (Drewes, 2010; Yamato, Monden, Matsumoto, Inoue, & Torii, 2000). According to Drewes (2010), it is difficult to achieve the same precision pointing with the eyes, as is possible with a computer mouse and it might therefore not be appropriate to replace a computer mouse with gaze control without a customized user interface. It has been suggested that the interface should be adjusted by enlarging buttons to make the eye control work effectively (Yamato et al., 2000). In this study, all the elements that the user could click on were enlarged. One part of the Communications Task was removed because the level of precision required was not manageable with gaze control. Two of the, authors who were not involved in implementing the modifications, were asked to test the interface with the eye-tracker to verify that the new interface could be used with gaze control. The same modified interface was used for all test cases regardless of interaction method.

The eyes are naturally used to search for information. The approach used in this study was that the main use for eye-gaze should be in visual search and selection tasks, while tasks like acknowledging and activating should be done using other modalities (Kumar, 2007). For the gaze control condition, the space bar on the keyboard was therefore used as a select or confirm button. This was also done to

avoid the “Midas-touch” problem, i.e. that the user would inadvertently interact with everything he/she looks at. MATB-II records several performance measures from the participants. The measures used in this study were: response time and correct responses in percent from the System Monitoring Task, deviation from the centre point in the Tracking Task, and the deviation from target fuel units (2500) in the Resource Management Task. Perceptual sensitivity, the discriminability index d’ (Green & Swets, 1974), was calculated based on the hits, misses, correct rejections and false alarms from the Communication Task.

Gaze control was implemented using an Eye Tribe device (Eye Tribe). This device consists of an eye-tracking bar mounted below a computer screen. The eye-tracker’s update frequency is 30-75 Hz, and the stated accuracy is 0.5-1° visual angle.

Procedure

The participants first completed an informed consent form and a demographics questionnaire. Next, an instruction video (ca. 8 minutes long) was used to introduce the participants to the different tasks of MATB-II. After the video, the instructor explained the NASA TLX form and provided a guide to use when completing the questionnaire later. The participants completed two training scenarios in MATB-II, the first time using the mouse to interact with the program and the second time using the gaze control equipment. The Eye Tribe equipment was calibrated and re-calibrated to the user before each gaze control scenario. A joystick was used for the tracking task in both the mouse and the gaze control conditions. The participants were encouraged to ask clarifying questions during the practice scenarios. After the practice scenarios, the participants completed four different scenarios:

- Gaze control with low task difficulty (GC-L).
- Gaze control with high task difficulty (GC-H).
- Mouse control with low task difficulty (M-L).
- Mouse control with high task difficulty (M-H).

The length of each scenario was five minutes. Task difficulty was manipulated by the length of the response windows and number of task interactions. The more difficult conditions requirered faster responses and more input from the participant. The response time for the scenarios was 20 s for the low task difficulty and 15 s for the high task difficulty. The order of the scenarios was blocked by control condition and counterbalanced across participants. The participants completed a NASA TLX questionnaire at the end of each scenario.

Results

Three participants failed to answer the NASA TLX within the 30-second time limit. This resulted in mental workload data from 27 participants and performance data from 30 participants.

Workload

The experienced mental workload was significantly higher with gaze control than with computer mouse, $F(1, 26) = 102.935$, $p < .001$ partial $\eta^2 = 0.798$ (see Figure 1). There was no significant difference between low and high task difficulty, $F(1, 26) = .294$, $p = .592$, partial $\eta^2 = 0.011$, and no significant task difficulty and control mode interaction, $F(1, 26) = .263$, $p = .613$ partial $\eta^2 = 0.010$.

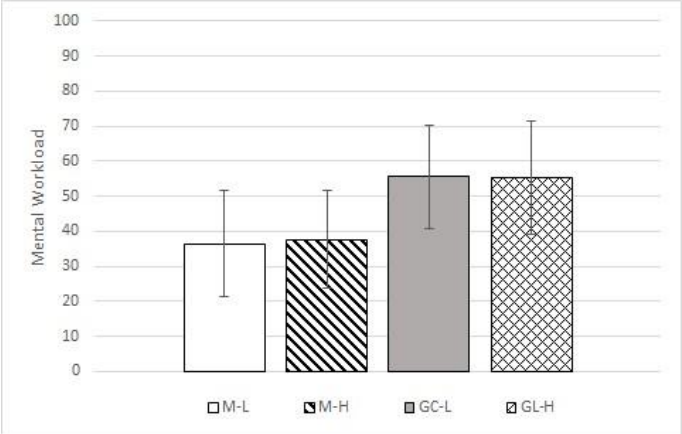


Figure 1. Mental workload as assessed with the total score of the NASA TLX for the different scenarios. M=Mouse control, GC= Gaze control, L = Low, H= High task difficulty

Performance

There was no significant difference between gaze control and computer mouse interaction for d'-scores in the communication task, $F(1, 29) = .214$, $p = .647$, partial $\eta^2 = 0.007$ (see Figure 2). The participants did however perform significantly better in the high task difficulty scenarios than the low task difficulty scenarios, $F(1, 29) = 63.076$, $p < .001$, partial $\eta^2 = 0.685$. There was no significant interaction between control mode and task difficulty, $F(1, 29) = 3.52$, $p = .558$, partial $\eta^2 = 0.012$.

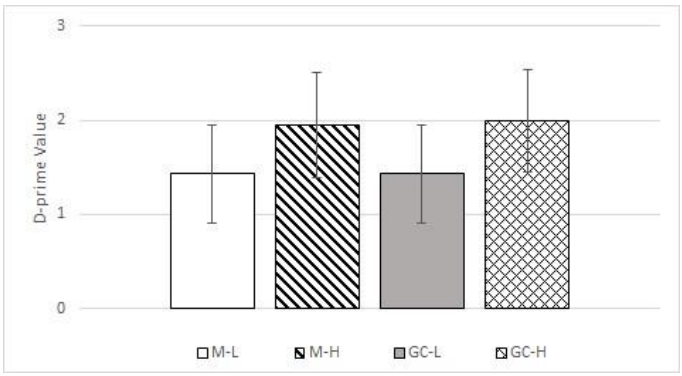


Figure 2. d' in the communications task. M=Mouse control, GC= Gaze control, L = Low, H= High task difficulty

There was no significant difference between gaze control and computer mouse interaction for fuel deviation in the resource management task, $F(1, 29) = .704$, $p = .408$, partial $\eta^2 = 0.024$, nor between high and low task difficulty, $F(1, 29) = 1.120$, $p = .299$, partial $\eta^2 = 0.037$, nor a control mode by task difficulty interaction, $F(1, 29) = .021$, $p = .886$ partial $\eta^2 = 0.001$.

For the system monitoring task the percentage of correct responses was significantly lower with gaze control than mouse control, $F(1, 29) = 11.768$, $p = .002$, partial $\eta^2 = 0.289$ (see Figure 3). However, there was no effect of task difficulty, $F(1, 29) = .494$, $p = .488$ partial $\eta^2 = 0.017$, nor a control mode by task difficulty interaction, $F(1, 29) = 0.019$, $p = .893$ partial $\eta^2 = 0.001$.

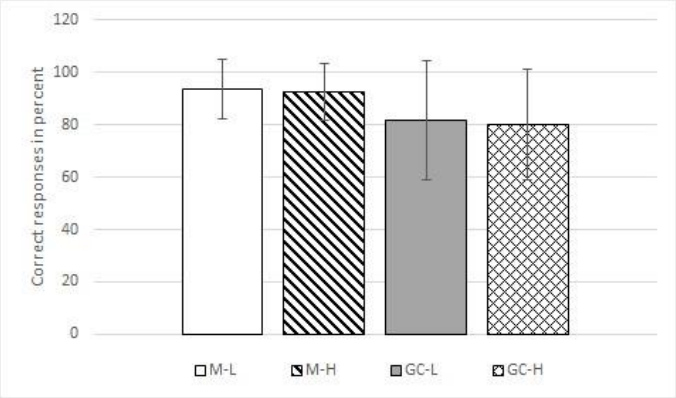


Figure 3. Correct responses in percent for the system monitoring task. M=Mouse control, GC= Gaze control, L = Low, H= High task difficulty

The response time for the system monitoring task was significantly higher for gaze control compared to mouse control, $F(1, 29) = 29.85$, $p < .001$, partial $\eta^2 = 0.507$ (see Figure 4). The response time was also significantly lower in the high task difficulty scenarios as compared to the low task difficulty scenarios, $F(1, 29) = 15.035$, $p = .001$, partial $\eta^2 = 0.341$. There was no interaction between control mode and task difficulty for the response time, $F(1, 29) = 1.967$, $p = .171$ partial $\eta^2 = 0.064$.

For the tracking task there was a significantly higher deviation from the centre for the gaze control mode than the mouse control mode, $F(1, 29) = 37.693$, $p < .001$, partial $\eta^2 = 0.565$ (see Figure 5). There was no effect of task difficulty, $F(1, 29) = 4.111$, $p = .052$ partial $\eta^2 = 0.124$, nor a control mode by task difficulty interaction, $F(1, 29) = .366$, $p = .550$ partial $\eta^2 = 0.012$.

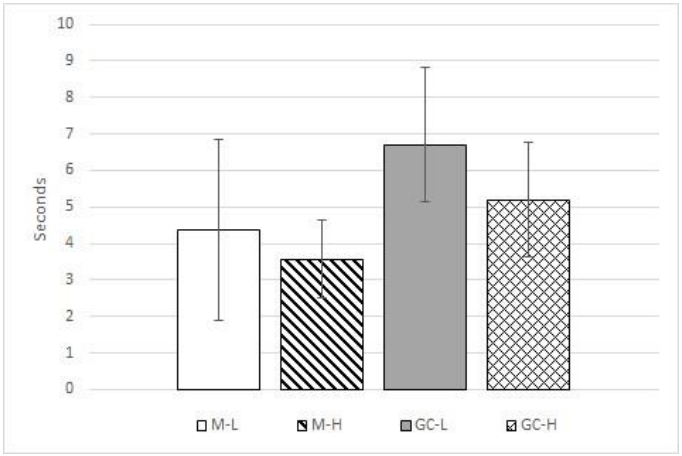


Figure 4. Response time in seconds for the system monitoring task. M=Mouse control, GC= Gaze control, L = Low, H= High task difficulty

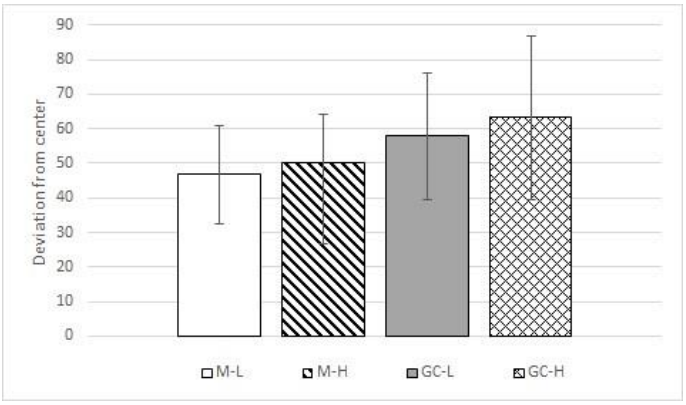


Figure 5. Deviation from centre for the tracking task. M=Mouse control, GC= Gaze control, L = Low, H= High task difficulty

Discussion

The results show that gaze control was associated with higher mental workload in all scenarios and decreased performance in two out of four scenarios in comparison with mouse control. These results indicate that gaze control was inappropriate for the tasks used in the current study. One reason could be that the eyes were used for both information retrieval and control, whereas the manual input was used only for control. Gaze control could therefore have led to resource conflicts (Wickens, 1980). Another reason could be that the interface was not optimally adapted for gaze control. Some adjustments with regard to the size of user control areas (e.g., buttons) were made but these were not subject to rigorous validation.

The participants did not rate the high task load scenarios as producing significantly higher workload, and performance was better for the high task load scenarios for the

auditory and system monitoring task. This could indicate that the experimental manipulation of taskload did not work as intended and did not increase the taskload appropriately. An alternative hypothesis as to why the response times were longer in the low task load scenarios than in the high task load scenarios could be a speed accuracy trade off (Wickens et al., 2012). When the demand for multitasking was lowered in the easy scenarios the participants could have focused more on the tasks demanding static input (i.e. resource management task and tracking task), as well as prioritizing precision over speed.

Regarding the objective to explore the use of a low cost environment for testing and research, the study shows promising results. The tested equipment and software turned out to be useful and a first step towards building a simulation environment for early evaluation of new concepts.

Future studies

Multiple resource theory (Wickens, 1980) describes how using different modalities may lower mental workload, and future studies could therefore be conducted to investigate the possibility of controlling some elements in MATB-II using gaze control and others using mouse control. This could explore the possibility of utilizing the advantages of gaze control complementary to the mouse control. Another modality to consider is vocal control, were voice commands could be used to perform clicks. This usage of a combination of different modalities might provide alternative action possibilities for the pilot in difficult situations.

The participants were all highly experienced in using computer mouse, which was expected, and none had previously used gaze control. Future studies could explore the effects of training and experience in using gaze control in simulated flight deck settings.

For a more complete view of the participant's mental workload, physiological measures could be added. This more quantitative approach may provide different insights than the subjective NASA TLX method that was used in this study. Some physiological measures of interest are pupil size and heart rate variability.

Conclusions

Gaze control was associated with higher mental workload in all scenarios and lower performance in two out of four scenarios. The different levels of task load affected task performance such that the participants performed better in the high task load scenarios. Overall, the results showed that the participants performed better and reported lower mental workload using a computer mouse as compared to gaze control in the simulated flight deck.

References

Aricò, P., Borghini, G., Graziani, I., Taya, F., Sun, Y., Bezerianos, A., Thakor, N., Cincotti, F., & Babiloni, F. (2014). Towards a multimodal bioelectrical framework for the online mental workload evaluation. In *36th Annual International Conference of the IEEE Engineering in Biology and Society*, Chicago, USA.

Charlton, S.G. (1996). Mental workload test and evaluation. In T.G.O´Brien & S.G. Charlton (Eds.), *Handbook of human factors testing and evaluation* (pp. 181-197). Mahwah, NJ: Lawrence Erlbaum and Associates.

Comstock Jr, J. R., & Arnegard, R. J. (1992). MATB - Multi-Attribute Task Battery for human operator workload and strategic behavior research *Technical Memorandum 104174*. Hampton, Virginia: NASA, Langley Research Center.

Drewes, H. (2010). *Eye gaze tracking for human computer interaction.* (Dissertation), LMU, München. (No. 11591)

Eggemeier, F.T., & Wilson, G.F. (1991). Performance-based and subjective assessment of workoad in multi-task environments. In D.L. Damos (Ed.), *Multiple-task performance* (pp. 217-278). London: Taylor & Francis.

Eye Tribe. (2016). The eye tribe tracker. Retrieved 2 February, 2016, from https://theEyeTribe.com/products/

Gopher, D., & Donchin, E. (1986). Workload-An Examination of the Concept. In K. R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of Perception and Human Performance* (Vol. 2). New York: Wiley.

Green, D.M., & Swets, J.A. (1974). *Signal detection theory and psychophysics*. Oxford, UK: Robert E Krieger.

Hancock, P.A., & Warm, J.S. (1989). A dynamic model of stress and sustained attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 31*, 519-537.

Hart, S.G., & Staveland, L.G. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press.

Hockey, G.R.J. (1997). Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework. *Biological Psychology, 45*, 73-93.

Jacob, R.J.K. (1991). The use of eye-movements in Human-Computer Interaction Techniques: What you Look at is What You Get. *ACM Transactions in Information Systems, 9*, 152-169.

Kumar, M. (2007). *Gaze-enhanced User Interface Design.* (Dissertation), Stanford University, Stanford.

NASA. (2014). MATB-II. *http://matb.larc.nasa.gov/*. Retrieved 2015-05-13

O´Connell, S.D., Castor, M., Pousette, J., & Krantz, M. (2012). *Eye tracking-based target designation in simulated close range air support.* In *the Human Factors and Ergonomics Society 56th Annual Meeting*, Boston, MA.

Penkar, A.M., Lutteroth, C., & Weber, G. (2013). *Eyes only: navigating hypertext with gaze.* Paper presented at the Human-Computer Interaction INTERACT, Cape Town, South Africa.

Rupp, M.A., Oppold, P., & McConnell, D.S. (2013). Comparing the performance, workload, and usability of a gamepad and joystick in a complex task. In *the Human Factors and Ergonomics Society 57th Annual Meeting*, San Diego, CA.

Santiago-Espada, Y., Myer, R. R., Latorella, K.A., & Comstock Jr, J.R. (2011). The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User´s Guide. Hampton, VA: NASA, Langley Research Center.

Svensson, E., Angelborg-Thanderz, M., Sjoberg, L., & Olsson, S. (1997). Information complexity-mental workload and performance in combat aircraft. *Ergonomics, 40*, 362-380.

Wickens, C.D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance VIII* (pp. 239-257). Hillsdale, NJ: Lawrence Erlbaum.

Wickens, C.D., Hollands, J.G., Banburry, S., & Parasuraman, R. (2012). *Engineering Psychology and Human Performance* (4th ed.). New York, NY: Pearson Education Inc.

Yamato, M., Monden, A., Matsumoto, K., Inoue, K., & Torii, K. (2000). Button selection for general GUIs using eye and hand together. In *The 5th International Working Conference on Advanced Visual Interfaces*, Palermo, Italy.

Eye activity measures as indicators of drone operators' workload and task completion strategies

Philippe Rauffet[1], Assaf Botzer[2], Alexandre Kostenko[1], Christine Chauvin[1], & Gilles Coppin[3]
[1]University of South Brittany, Lorient, France
[2]Ariel University, Israel
[3]Telecom Bretagne, Brest, France

Abstract

We studied whether eye activity patterns in a simulated drone-operating task could be associated with workload levels and task completion strategies. Participants sent drones to suspected areas according to messages they received and according to self-initiated search. They were also required to validate whether suspected targets were indeed hostile prior to attacking them. We tested whether the number of suspected targets affected the number of eye transitions between task zones and whether it affected fixation duration in different task zones. We found that operators made fewer transitions between task zones as the number of targets increased. This was because they focused more on one zone and not on the others. Interestingly, the zone they attended to relatively more was the one they needed for attacking targets and not the ones where targets usually appeared. This was probably because attacking required extended cognitive operations. Findings demonstrated that eye activity patterns can be used to infer task completion strategies and to identify workload levels, once these strategies are described. Workload levels and task completion strategies should therefore be studied by a combination of hypothesis driven and exploratory driven methods. Eye activity patterns can then be used as triggers for assisting overloaded operators.

Introduction

Mental workload (MWL) and task completion strategies are interrelated. They both affect the degree of successful task completion as operators change their strategies and implement different mental workload's regulation loops for improving task performance or decreasing task induced cognitive load (Hockey, 1997; Kostenko, Rauffet, Chauvin, & Coppin, 2016; Schulte, Donath, & Honecker, 2015). Designers of work environments would therefore like to gain insight into the levels of MWL that human operators experience and to learn about their task completion strategies. Such insights may assist in integrating adaptive automation to support operators in times of high MWL. As design validation criteria, These ocular MWL indicators can also point out the design of harder-to-operate interface components and may inform training programs to correct less productive task completion strategies (Byrne & Parasuraman, 1996; Hockey, 1997; Nickel & Nachreiner, 2003; Van Orden et al., 2001).

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Performance measures, self-report measures (e.g., NASA-TLX) and physiological measures (e.g., heart rate variability; skin conductance) are all widely acceptable for estimating workload and to some extent for learning about task completion strategies. At the same time, however, they often fall short in detecting short periods of elevated workload that may trigger changes in task completion strategies, or conversely, indicate the recent occurrence of such changes (Hilburn & Jorna, 2001; Verwey & Veltman, 1996). Thus, these measures may sometimes not provide the information that practitioners would like to have. Eye activity measures, in contrast, are more adapt for learning about short periods of elevated workload and about shifts in task completion strategies. This is because they indicate what sources of visual input operators attend to and for how long, and they are therefore indicative of the cognitive processes that this input serves (Rayner, 1998; Salvucci, 2001).

For instance, Bijleveld, Custers & Aarts (2009) demonstrated that greater potential rewards in a digit-retention task led to increases in pupil dilation as would be expected when people invest more effort in tasks. Salvucci (2006) demonstrated how cognitive modelling of driver lane keeping, curve negotiation and lane changing corresponded with gaze distribution in driving. Botzer et al. (2015) demonstrated that aid from automation in a simulated quality control task led to changes in search patterns of faulty items and to changes in how much time decision makers inspected items. These changes corresponded with decision makers reported effort. Finally, Van Orden et al. (2001) demonstrated that blink frequency, fixation frequency and pupil diameter could be used to predict fluctuations in target density in a simulated anti-air-warfare task.

Still, in Van Orden et al. (2001), gaze distribution patterns of some participants were not related to task completion strategies, but rather to task disengagement. Next, while higher workload led to greater frequency of fixations but not to changes in fixation durations in a visual search task (Zelinsky et al., 1997), higher cognitive processing load did correspond with longer fixation durations in a flight task (Callan, 1998). Thus, MWL affects eye activity in different ways depending on the task, and one should therefore interpret eye activity measures according to task characteristics and according to prior knowledge and hypotheses. Exploratory inspection of the data together with alternative hypotheses about operators' cognitive processes should be also employed to learn about task completion strategies and operators' MWL.

Based on the methods and insights from a body of research on eye activity measures and human performance, we set to explore drone operators' MWL and task completion strategies. We expected that greater density of hostile targets that operators need to handle in a simulated drone-operating task would lead to greater fixation durations in certain task zones and to fewer gaze transitions between task zones.

Method

Participants

Twenty-two participants, aged from 18 to 20 (mean: 19, standard deviation: 0.7) took part in the experiment. For reasons of homogeneity, all were men and had good experience with video gaming.

Simulation set-up and experimental task

The task was a simulation of securing an area with a swarm of drones that we ran on a demonstrator, named SUSIE (Coppin & Legras, 2012). SUSIE is supported by Java software, and it allows participants to interact with and to supervise a swarm of drones using a mouse-screen system.

Only one operator is required, but some tasks can be or are achieved by an artificial agent. The system provides different information to the operator from two sources: a dynamic map and a message banner (Figure 1). The dynamic map gives information about the areas that the drones control such as the vehicles in these areas and their state. The message banner indicates the coordinates and direction of a vehicle that the operators need to assign high priority to its neutralization.

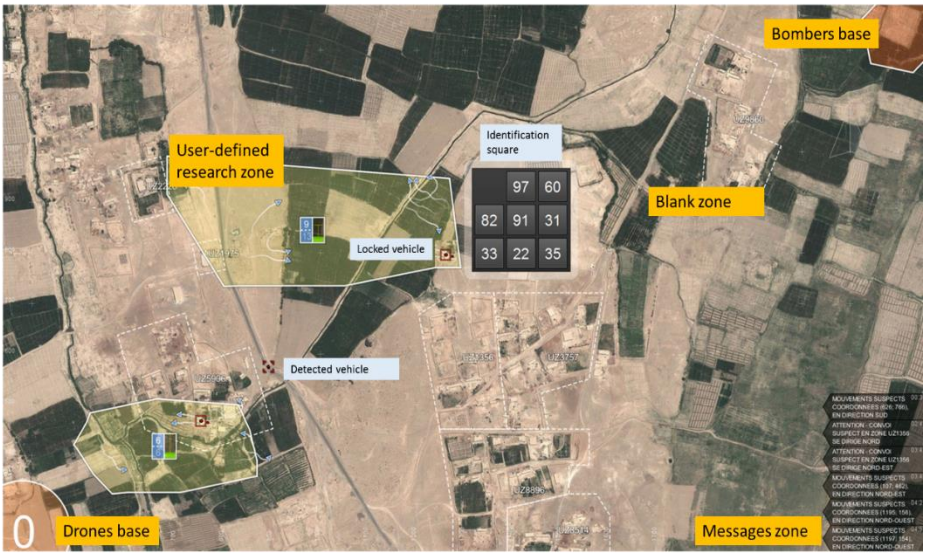


Figure 1. Dynamic monitoring map

The main task is to detect and neutralize the threats (i.e., hostile vehicles) on the map. When a vehicle is generated by the software, it is hidden, i.e. it is present on the map but invisible (it has to be detected by drones sent by the participant). Before it is neutralized, the status of the vehicle changes several times (Fig. 2).

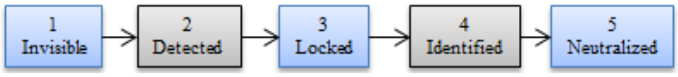


Figure 2. The different vehicle states

To advance from one status to the next, operators need to complete a number of sub-tasks, related to different areas of interest (AOI). Part of these AOIs, as the message zone, drone base and bombers base, are static and part, as the user-defined research zones and the blank zone, whose surface changes according to the creation or removal of user-defined research zones, are dynamic. Note that blank zones are used to identify vehicles, but also to define new user-defined research areas. In other words, the blank zone is all the space of the map where there are no specific AOIs. It is on this blank zone that participants create new user-defined research zones, and it is on this blank zone that participants achieve the identification sub-task (Table 1).

Table 1. Synthesis of the tasks and the associated AOI

Tasks	**Description**	**Associated Areas Of Interest (AOI)**
T1: Read a message to locate a vehicle	Extract information on a suspected vehicle (coordinates and direction) from a message.	**Message zone**
T2: search for a new vehicle	Search vehicle by creating a user-defined research zone. The drone moves automatically in a zone, a vehicle is detected when a drone is flying over it.	**Blank zone** (participants have to create new research zones on the map)
T3: Lock a detected vehicle	Lock a vehicle (from state 2 to state 3) by flying over it a second time.	**User-defined research zones**
T4: Identify a vehicle as a target	Select and identify a vehicle as a threat or not (from state 3 to state 4). In the experimental version, this task is simulated by a basic cognitive task, which consists of sorting out nine numbers in an increasing order.	**User-defined research zones** (participants select a locked vehicle in a research zone) **Blank zone** (participants make the identification on a number square, appearing on the map, outside the research zones)
T5: Attack a target	Neutralize a vehicle (from state 4 to state 5) by drawing a corridor starting from the bombers base, which simulates the sending of a helicopter.	**Bombers base**
T6: Manage drones in user-defined research zones	Regulate the number of drones in a research zone. Drones go back to the drones' base when fuel tank is empty, and participants have to refill research zones	**Drones base**

	with drones.	

Performance of the sub-tasks in Table 1 was subject to temporal constraints:

- A detected vehicle had to be locked within 5 seconds following detection. Otherwise, it had to be detected once again.
- A locked vehicle had to be selected within 10 seconds following locking. Otherwise, it turned automatically from "locked" to "detected".
- An identified vehicle had to be neutralized within 100 seconds following its identification. Otherwise, it turned automatically from "identified" to "detected".

Scenario and objectives

The experimental session lasted 20 minutes and had two phases: Phase A of low difficulty (10 minutes) and phase B of high difficulty (10 minutes). We controlled the difficulty level by changing the rate of new vehicles and new messages. To limit the effects of the order in which the phases were presented (effect of learning, etc.) two similarly sized groups of participants were created. The first one performed phase A first, and the second group performed phase B first.

We communicated two objectives to all participants:

- *Reaching a minimum number of neutralized targets:* participants have to neutralize more than 25 vehicles during the mission.
- *Considering all priority targets:* For all messages (giving the location of priority targets), participants have to draw a zone according to the message in no longer than two minutes.

Data collection and processing

Measures and sensors

The system supporting SUSIE software is composed of a screen of 24" and a mouse connected to a PC. The devices used for data acquisition were an eye tracker FaceLAB5© for pupillary response, the log (text file) of scenario events (vehicle appearance, messages) and operator's mouse actions recorded on SUSIE.

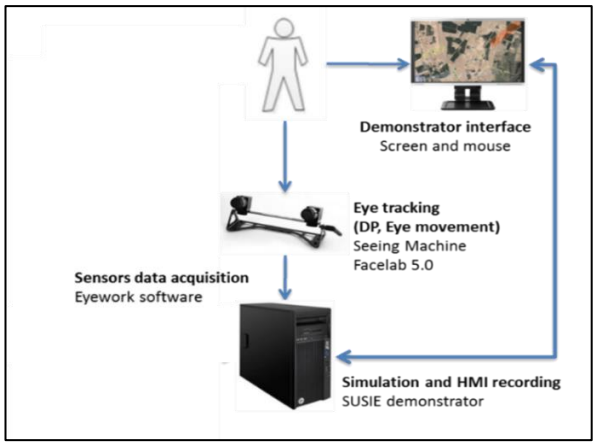


Figure 3: Experimental materials

Using SUSIE's log, one is able to compute indicators related to the density of stimuli and to operator performance:

- *Density level:* this variable characterizes the real-time task constraint, due to the density of informational stimuli relevant to dealing with the two objectives given to the participants. We defined density as the sum total of targets and messages that operators have not yet processed. This density varied from 0 to 35 stimuli across all participants and all experimental sessions, and we then categorized it into three density levels (low level: lower than 7 stimuli, medium level: between 8 and 15 stimuli, high level: higher than 16 stimuli).
- *Performance on message processing:* This is a binary indicator, computed every second. It decreases to 0 as soon as a message is not processed in time (i.e. if a new research zone is not created around the coordinates in the message during the 2 minutes following its appearance), and it stays or increases to 1 otherwise.
- *Performance on target neutralization:* This is a binary indicator, computed every second. It decreases to 0 as soon as a target is not neutralized fast enough (i.e. if time from first detection exceeds 2 minutes and target had still not been neutralized), and it increases to 1 otherwise.
- *Global Performance:* this indicator was computed as the mean of the two previous indicators.

In parallel to performance data, we also extracted eye movement variables that were related to task completion strategies using the Facelab 5.0 eye-tracking system:

- *Horizontal and vertical gaze concentration:* We defined this variable as the standard deviation in pixels (px) of gaze position on the X and Y axes (Wang, Reimer, Dobres & Mehler, 2014).
- *Proportion of Total Fixation Duration on each AOI:* It corresponds to the percentage of time spent to consult a specific AOI, based on fixation durations, as described in Masthoff, Mobasher, Desmarais & Nkambou (2012, p.132)
- *Transition rate between each pair of AOI:* We defined this variable as the number of gaze transitions per second from one AOI to another (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka & Van de Weijer, 2011, p. 424).

Density level was used to segment the experimental data over time. Each change in density level results in the ending of a temporal segment and the creation of a new one. The variables related to performance and eye activity were averaged on these segments.

Experimental protocol: training and briefing

The experiment was conducted in individual sessions of approximately 2 hours at the Lab-STICC laboratory and was divided into six phases:

1. Greeting participants: completing the profile questionnaire (age, video game experience)

2. Explaining SUSIE principles: a slide presentation was used to explain (verbally) the monitoring and drone management tasks and the interface. The experimenter provided the objectives of the mission that participants performed on the simulator.

3. Practicing and training: participants carried out the tasks to familiarize themselves with drone-swarm based monitoring activities.

4. Sensors parameters and calibration setting: Eye-Works Record software was launched to record the participants' eye movements after faceLAB software had calibrated head, eyes, and test environment;

5. Carrying out the monitoring tasks: participants carried out the twenty-minute task scenario on SUSIE and completed a questionnaire at the end of the experiment;

6. Debriefing and thanking: We asked participants for their view of the set of proposed tasks, the interface, and the simulator in order to obtain their comments. To ensure inter-subject independence of the collected data, participants were asked not to share test contents with those around them.

Experimental design and hypotheses

We analysed the data from 17 participants after excluding 5 participants from the analysis due to intermittent failures in eye activity data acquisition. All were subject to the same scenario of twenty minutes, during which different constraint levels occurred. Our experimental design was thus a mixed factorial design of 17 Participants x 3 Constraint Levels. Table 2 presents the independent variables (related to informational constraint), and the dependent variables (related to performance and eye movement strategies). Statistical analyses of the data were conducted according to a General Linear Model.

Table 2: Independent and dependent variables

Independent variable	**Indicator of**	**Range**
Constraint Level	Informational density	3 ordinal classes: Low, Medium, High
Dependent variables	**Indicator of**	**Range**

Performance on message processing	Performance	Continuous variable, from 0 to 1 (mean=0.462; std=0.493)
Performance on target neutralization	Performance	Continuous variable, from 0 to 1 (mean=0.523; std=0.498)
Global performance	Performance	Continuous variable, from 0 to 1 (mean=0.493; std=0.351) $\left(\frac{\sum_{i=1}^{N} STC_i}{N}\right)$ Where STCi is whether a subtask is completed or not when the sample is taken (yes=1 no=0) and N is the number of samples in the considered period
Gaze concentration	Eye movement strategies	***Horizontal concentration:*** continuous variable, from 0 to 852 (mean=323.7; std=171.1) ***Vertical concentration:*** continuous variable, from 0 to 537 (mean=210.9; std=104.2)
Proportion of total fixation durations on each AOI	Eye movement strategies	***Message zone:*** continuous variable, from 0 to 100 (mean=17.2; std=26.7) ***Drones base:*** continuous variable, from 0 to 100(mean=10.0; std=19.9) ***Bombers base:*** continuous variable, from 0 to 100(mean=5.5; std=16.8) ***User-defined research zone:*** continuous variable, from 0 to 100(mean=58.0; std=36.7) ***Blank zone:*** continuous variable, from 0 to 100(mean=9.3; std=23.1)
Transition rate between AOI	Eye movement strategies	***Number of transitions/s between all AOIs:*** continuous variable, from 0 to 5.22 (mean=0.60; std=0.65) ***Number of transitions/s between each pair of A0I:*** Every 2-AOI transitions were also analysed in terms of frequency

We posit the following five hypotheses with respect to the effects of constraint level on participants' eye movement strategies and performance.

Effect of density level on eye movement strategies
Higher density of informational stimuli should result in attentional funnelling or perseveration on some task-related areas as a means to accommodate task demands.

We therefore hypothesize that:

- H1: Gaze concentration will decrease when density of stimuli (messages and targets) increases.
- H2: Mean transition rate between all AOIs will decrease when the density of stimuli increases.
- H3: Proportion of total fixation durations on some AOIs will increase when the density of stimuli increases.
- H4: Transition rates between certain pairs of AOIs would increase when density of stimuli increases (this is because participants would give higher priority to some task areas and not to others).

Effect of density level on performance

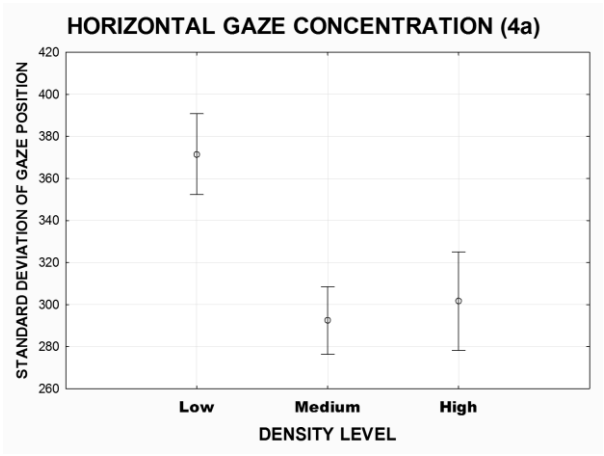
- H5: We expect that a higher density of targets and messages will result in performance decrements.

Results

Effect of density level on eye movements

To evaluate the global gaze behaviour of participants we analysed their gaze variability and transition rate between AOIs. This context-independent analysis resulted in two main observations:

- H1 was partially supported, as we found a significant effect of the density of stimuli on horizontal gaze variability ($F(2,901)=20.657$, $p<.001$), but no significant effect on vertical gaze variability. As depicted in figure 4a, the gaze is more concentrated in the medium and high density levels (M=294px and M=301px, respectively) than in the low density level (M=373px). With respect to the lack of effect of density level on vertical gaze concentration, we believe that the reason for this may be that most activity remained in a large central, vertical strip on the display, whether density was lower or higher. This will be demonstrated in our analyses of fixation durations and gaze transitions.
- The effect of density on gaze concentration corresponded with the results from the analysis of transition rate between all AOIs (cf. figure 4b) and corresponded with H2. Overall, participants made fewer transitions per second between AOIs when the density of stimuli increased ($F(2,946)=5.783$, $p<.005$), dropping from M=0.72 transitions/sec in low density level to M=0.51 transitions/sec in the high density level.



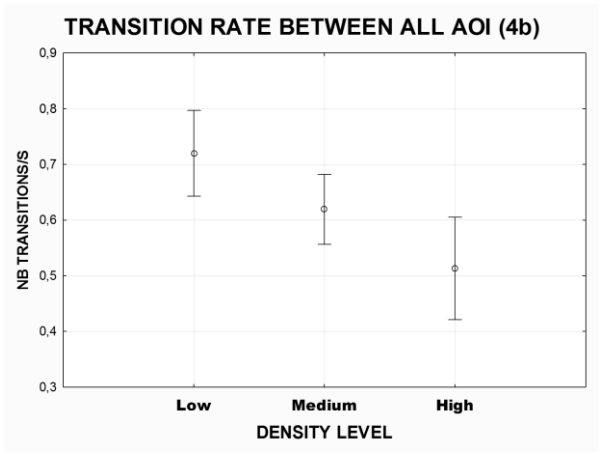
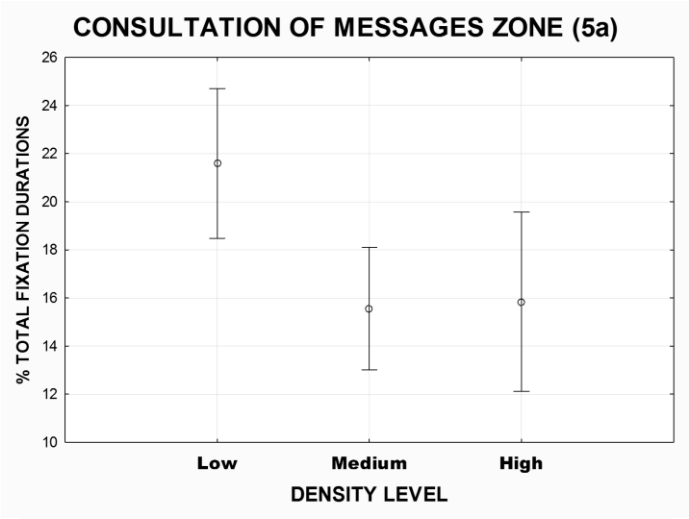


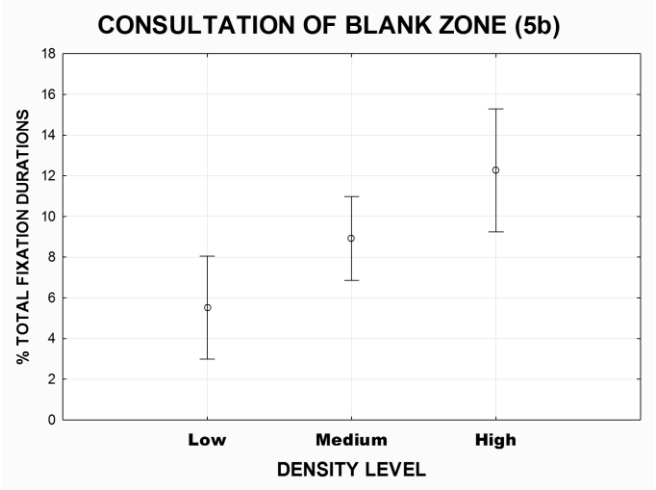
Figure 4: Horizontal gaze concentration (4a) and transition rate between AOI (4b). Vertical bars denote 0.95 confidence intervals.

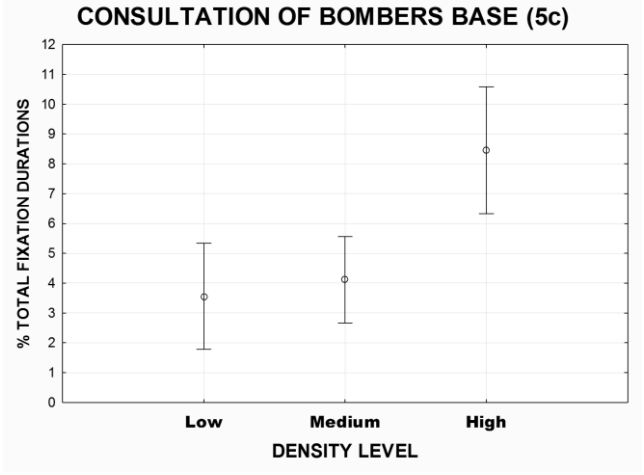
To evaluate the eye movement patterns within task context we studied the proportion of total fixation durations on each AOI and the rate of transitions between each pair of AOIs of the five areas of interest we described in section 2.2 in the Method section.

Figure 5 shows a number of significant effects of the density level on fixation distribution and duration on AOIs. The statistical results highlighted three main observations:

- Participants focused relatively less on message zones when the density level increased ($F(2,946)=4.84$, $p<.01$). Proportion of total fixation durations dropped from M=21.8% in low density level to M=15.7% and M=15.9% for medium and high density levels, respectively (cf. figure 5a).


CONSULTATION OF MESSAGES ZONE (5a)
% TOTAL FIXATION DURATIONS
26
24
22
20
18
16
14
12
10
Low
Medium
High
DENSITY LEVEL


CONSULTATION OF BLANK ZONE (5b)
% TOTAL FIXATION DURATIONS
18
16
14
12
10
8
6
4
2
0
Low
Medium
High
DENSITY LEVEL


CONSULTATION OF BOMBERS BASE (5c)
% TOTAL FIXATION DURATIONS
12
11
10
9
8
7
6
5
4
3
2
1
0
Low
Medium
High
DENSITY LEVEL

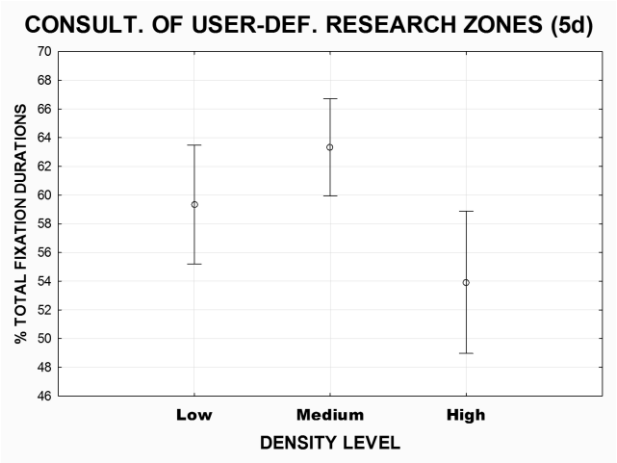


Figure 5: Proportion of total fixations durations on messages zone (5a), blank zone (5b), bombers base (5c) and user-defined research zones (5d). Vertical bars denote 0.95 confidence intervals.

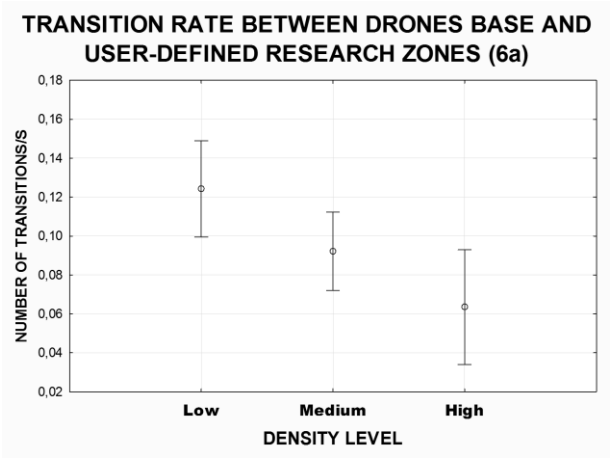
- Consultation of blank zones - associated with identification of vehicles and consultation of bombers base associated with the neutralization of vehicles increased with density level ($F(2,946) = 5.74$, $p<0.005$ and $F(2,946) = 7.05$, $p<.001$, respectively). The proportion of total fixation durations increased from M=5.8% (low level) to M=12.1% (high density level) in blank zone, and from M=3.6% (low level) to M=8.4% (high density level) in bombers base (cf. figures 5b and 5c).
- Finally, participants focused relatively less on user-defined research zones (cf. figure 5d) when density level was higher ($F(2,946) = 4.82$, $p<0.001$) (M=54%, M=59.5% and M=63.7% for high, medium and low density levels, respectively).

Thus, in accordance with H3, higher density of stimuli led participants to focus more on certain task areas and less on other, with one exclusion - we found no significant effect of the density level on the distribution of fixations on the drones' base. This is probably because neither when the density of stimuli was lower nor higher, was the monitoring of drone state a highly frequent sub-task.

Findings from the analyses above corresponded with a frequency analysis of AOI consultation and corresponded with H4:

- Transition rate between drones base and user-defined research zones decreased when density level increased ($F(2,907) = 4.936$, $p<.01$), varying from M=0.123 transitions/s in low density to M=0.062 transitions/sec in high density.
- Participants executed less transitions between message zone and user-defined research zones when density level was high ($F(2,907) = 9.45$, $p>0.001$). The number of transitions per second varied from M=0.121 (low level) to M=0.060 (high level).

No significant effect of the density level on transitions between the other pairs of AOIs was found.



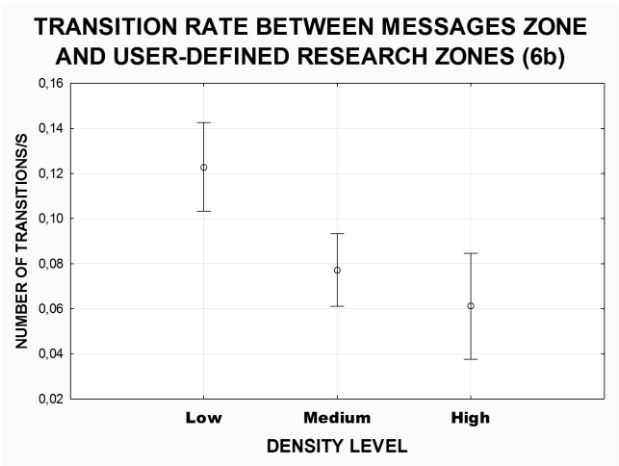


Figure 6: Transition rate between drones base and user-defined research zones (6a), and between messages zone and user-defined research zones (6b). Vertical bars denote 0.95 confidence intervals.

3.2. Effects of density level on performance

Different indicators were analysed: participants' performance on message processing (figure 7b), target neutralization (figure 7c) and on the sum total of the two former objectives (figure 7c). two main observations were made:

- In accordance with H5, density level had a significant effect on global performance ($F(2,946)=8.532$, $p<.001$). Global performance decreased when density level increased ($M=0.55$, $M=0.46$ and $M=0.44$ in low, medium and high density levels, respectively).

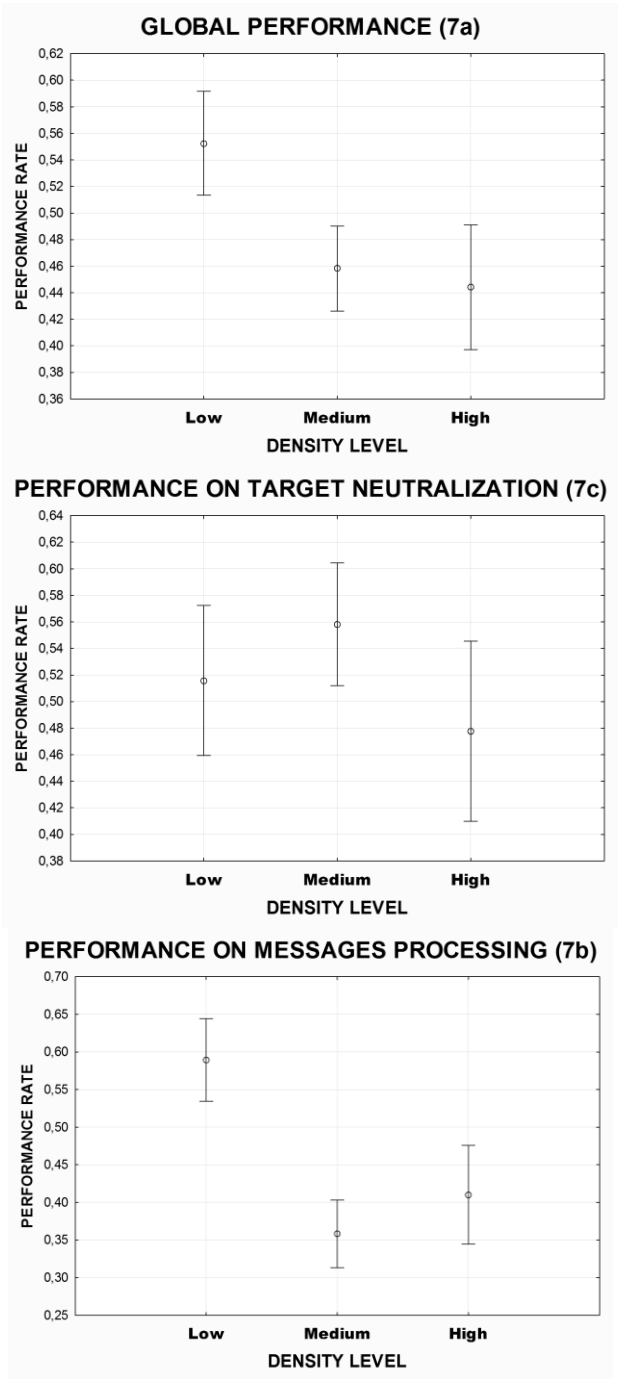


Figure 7: Performance rate for both message and target processing (7a), for message processing (7b) and for target neutralization (7c). Vertical bars denote 0.95 confidence intervals.

- This decrease of global performance is mainly explained by a decrease of performance on message processing ($F(2,946)=21.13$, $p<.001$) ($M=0.59$, $M=0.35$ and $M=0.41$ for low, medium and high density levels, respectively). To the contrary, no significant effect was found on the performance on target neutralization ($F(2,946)=1.976$, $p=.139$).

Discussion

Our analyses of operators' eye activity patterns and performance supported the main hypotheses. It was found that when the density of targets increased the overall variance in gaze positions along the X-axis was lower for medium and high compared to lower target density levels (Figure 4). This finding corresponded with a finer grained analysis showing that operators made fewer transitions between display areas and concentrated for relatively longer periods on certain areas and not on others. It was also found that performance was lower when the density of targets increased. This finding suggested that MWL was able to be manipulated and that changes in gaze behaviour were probably associated with different task completion strategies in response to changes in MWL. A deeper look into the eye activity patterns in terms of task dependent areas revealed how operators chose to deal with higher levels of MWL. When target density increased, operators reduced the proportion of time they spent on inspecting messages and by extension, the proportion of time they spent on inspecting self-defined research zones that one defines based on the messages. In contrast, operators increased the proportion of time that they spent on inspecting blank zones and the bombers' base that they needed for eliminating targets (Figure 5a and 5c). Performance measures corresponded with the strategy that eye activity measures reflected as greater number of targets led operators to losing points for not dealing with messages but not for failing to eliminate the targets that they had already detected (Figure 7).

The findings demonstrated the importance of context dependent interpretation of eye activity measures. It could be hypothesized that operators would invest different times in display areas when confronted with different levels of MWL, yet it was difficult to detail what areas they would inspect more or less carefully. The rate that operators could lose points for not dealing with messages and for not eliminating the detected targets was similar. Either way, score would drop down to zero every two minutes without action. Operators opted to prioritize the elimination of targets, probably because it made more sense to do so given the context of the task (securing a perimeter). In future studies, we will explore whether eye activity patterns would be sensitive enough to detect different strategies operators might employ in response to different rates of losing points (e.g., every 20 s for not dealing with messages and every two minutes for not eliminating the detected targets).

Eye activity measures corresponded with performance measures, showing their validity as a means to learning about task completion strategies in response to changes in MWL. At the same time, the correspondence between measures might raise the question why using eye activity measures in the first place and not just inferring about different levels of MWL and their consequences from operator performance? In this respect performance may sometimes remain unchanged even

when MWL increases, if people succeed to counter higher task demands by investing more effort. It is thus necessary to use a combination of measures when investigating MWL (Yeh & Wickens, 1988). Further, we computed average scores across 22 participants to detect a drop in performance between the three target density levels that we defined in the Method section. However, performance of individual operators may drop in response to lower/higher densities than the ones we defined but it would be impossible to anticipate this. Eye activity patterns, on the other hand, may be used as precursors of a later drop in performance. In future studies, with larger number of participants, we intend to test whether the changes that we described in the relative time operators spend in different display areas can indeed anticipate a later drop in performance. If this is indeed the case, then eye activity patterns may possibly serve as the basis of on-line algorithms to detect short periods of elevated MWL of single operators and trigger automatic or human assistance for those that experience too high task demands.

5. Conclusions

Different eye activity patterns can be detected in response to different levels of target density in a simulated drone-operating task. These patterns corresponded with operators' scores in the task, suggesting that eye activity measures can be used to detect short periods of elevated MWL and changes in task completion strategies. The findings, therefore, constitute a platform for further investigation of the practical usage of eye activity measures in work environment. In the future, we intend to investigate the sensitivity of the indicators of MWL that we described for detecting short periods of elevated MWL of individual operators as a means to triggering automatic or human assistance in tasks.

References

Bijleveld, E., Custers, R., & Aarts, H. (2009). The unconscious eye opener pupil dilation reveals strategic recruitment of resources upon presentation of subliminal reward cues. *Psychological Science*, *20*, 1313-1315.

Botzer, A., Meyer, J., Borowsky, A., Gdalyahu, I., & Shalom, Y. B. (2015). Effects of cues on target search behavior. *Journal of Experimental Psychology: Applied*, *21*, 73-88.

Byrne, E.A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology*,*42*, 249-268.

Callan, D.J. (1998). Eye movement relationships to excessive performance error in aviation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 42, No. 15, pp. 1132-1136). Chicago, Illinois: SAGE Publications.

Hilburn, B., & Jorna, P.G.A.M. (2001). Workload and air traffic control. In P.A. Hancock, & P.A. Desmond (Eds.), *Stress, workload, and fatigue*. Mahwah, NJ: L. Erlbaum.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford, UK, Oxford University Press.

Hockey, G.R.J. (1997). Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework. *Biological Psychology, 45*, 73–93.

Kostenko, A., Rauffet, P., Chauvin, C., & Coppin, G. (2016, August). A dynamic closed-looped and multidimensional model for Mental Workload evaluation. In *13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (HMS).* Kyoto, Japan

Masthoff, J., Mobasher, B., Desmarais, M., & Nkambou, R. (Eds.). (2012). User Modeling, Adaptation, and Personalization: *20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012 Proceedings*. Springer.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors*, *45*, 575-590.

Rayner, K. (1998). Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin, 124*, 372-422.

Salvucci, D.D. (2001). An integrated model of eye movements and visual encoding. *Journal of Cognitive Systems Research, 1*, 201-220.

Salvucci, D.D. (2006). Modeling driver behaviour in a cognitive architecture. *Human Factors*, *48*, 362-380.

Schulte, A., Donath, D., & Honecker, F. (2015). Human-System Interaction Analysis for Military Pilot Activity and Mental Workload Determination. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2015 (pp. 1375-1380).

Van Orden, K.F., Limbert, W., Makeig, S., & Jung, T.P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, *43*, 111-121.

Verwey, W.B., & Veltman, H.A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology: Applied*, *2*, 270-284.

Wang, Y., Reimer, B., Dobres, J., & Mehler, B. (2014). The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transportation Research Part F: Traffic Psychology and Behaviour*, 26, 227-237.

Yeh, Y.Y., & Wickens, C.D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30, 111–120.

Zelinsky, G.J., Rao, R.P., Hayhoe, M.M., & Ballard, D.H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, 448-453.

A method for quantitative estimate of risk probability in use risk assessment

Monica Tavanti[1] *& Lee Wood*[2]
[1]*Hoffmann-La Roche,* [2]*Medical Human Factors AG*
Basel, Switzerland

Abstract

Risks analysis requires that foreseeable risks related to usage of medical products are assessed and that mitigations are in place, so that use related hazards are as low as reasonably possible. It is commonly understood that risk is defined as the combination of the probability of occurrence of harm and the severity of that harm. International standards (notably ISO14971-2012) suggest that "foreseeable sequences of events that can produce hazardous situations and harm" should be considered, since: "a hazard cannot result in harm until […] a sequence of events […] lead to a hazardous situation". Sequence of events is probabilistic, and should be described considering the probability of hazards leading to hazardous situations (P_1), combing with the probability of hazardous situations leading to harms (P_2). P_1 and P_2 are essential in that their joint probability defines the likelihood of occurrence of harm, which is defined as, the "physical injury or damage to the health of people". Whilst the international standards suggest that use errors have to be considered, it does not clarify how to connect the probability of hazardous situations occurring (P_1) and the probability of hazardous situation leading to harm (P_2) with the probability that users may commit those errors. The present work proposes a method which enables quantitative estimations of use errors and clearly relates them to P_1 and P_2 estimates.

Introduction

This work relates to the assessment of use related risk in the use of Medical Devices and Drug/Biologic-Device Combination Products. Medical devices can be broadly defined as apparatuses intended for the diagnosis or treatment of disease.

The regulatory definitions of drug/ biologic-device combination products can vary based on the region. For the purpose of this article, the US 21CFR part 3.2 (e) definition shall apply, which defines a combination product as:

> (1) A product comprised of two or more regulated components, i.e., drug/device, biologic/device, drug/biologic, or drug/device/biologic, that are physically, chemically, or otherwise combined or mixed and produced as a single entity;

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

> (2) Two or more separate products packaged together in a single package or as a unit and comprised of drug and device products, device and biological products, or biological and drug products;
>
> (3) A drug, device, or biological product packaged separately that according to its investigational plan or proposed labeling is intended for use only with an approved individually specified drug, device, or biological product where both are required to achieve the intended use, indication, or effect and where upon approval of the proposed product the labeling of the approved product would need to be changed, e.g., to reflect a change in intended use, dosage form, strength, route of administration, or significant change in dose; or
>
> (4) Any investigational drug, device, or biological product packaged separately that according to its proposed labeling is for use only with another individually specified investigational drug, device, or biological product where both are required to achieve the intended use, indication, or effect.

The peculiarity of such devices used for the delivery of medicines resides in the fact that risks pertain two main components of the product: 1) the device(s), with any potential usability issues associated with their intended use or reasonably foreseeable misuse, and; 2) the medicinal product, whose administration entails potential risks associated with medication error, for example, administration by the wrong route, wrong dose or wrong rate (Vincent, 2010). When use errors occur, users are potentially exposed to hazards, hazardous situations, and, ultimately, to harm.

For Vincent (2010), the role of harm prevention is central to patient safety, and more prominent than prevention of error. Focusing on patient safety requires the manufacturers of devices and combination products to implement risk management processes.

Requirements for use related risk management

Medical device and medicinal product manufacturers are expected to perform risk analysis to identify known and foreseeable use related risks and to mitigate these risks to be within acceptable limits (WHO, 2016; EMA, 2015a, FDA, 2006).

The requirement to analyse and evaluate risks associated with use is broadly represented across regulation, harmonized/ recognized normative technical standards and a range of Health Authority guidance. The following section summarizes the current regulatory basis:

Under US Medical Device Quality System Regulation; risk analysis is a requirement of 21CFR820.30 (g) Design Validation (also applicable to combination products in accordance with 21CFR part 4):

> *"...Design validation shall ensure that devices conform to defined user needs and intended uses and shall include testing of production units under*

quantitative estimate of risk probability

> *actual or simulated use conditions. Design validation shall include software validation and risk analysis, where appropriate*…."

In the European Union (EU), the Essential Requirements of the European Medical Device Directive herein referred to as the MDD (Council Directive 93/42/EEC of 14 June 1993, incl. 2007/47/EC), states that the devices must be designed and manufactured in such a way as:

> *"…reducing, as far as possible, the risk of use error due to the ergonomic features of the device and the environment in which the device is intended to be used (design for patient safety)".*

The requirement for use related risk analysis is also integral to two Medical Device normative standards as harmonized to the EU MDD, and recognized as Medical Device consensus standards by the US Food and Drug Administration (FDA), namely:

- IEC 62366-1:2015 Medical Devices - Part 1: Application of Usability Engineering to Medical Devices, is recognized by the US FDA, currently an earlier version of the standard EN 62366:2008 is harmonized to the EU MDD.

- EN ISO 14971:2012 Medical devices - Application of risk management to medical devices, is a harmonized standard to the EU MDD, currently the preceding version of the standard ISO 14971:2007 (R) 2010 is a recognized consensus standard by the US FDA.

Finally, the importance of use related risk management as part of the Human Factors (HF) process for the development of medical devices and drug-device combination products is also underscored by a wealth of guidance from the US Food and Drug Administration (FDA 2016a, 2016b, 2016c), European Medicines Agency (EMA 2015b), the UK`s Medical and Healthcare Products Regulatory Agency (MHRA, 2016), and importantly the International Conference of Harmonization of technical requirements for registration of pharmaceuticals for Human use guidance: ICH-Q9 (EMA, 2015a; FDA 2006).

Integration of Risk Management into the HF Engineering process

Use related risk management is integral to the HF Engineering/ Usability Engineering (UE) process as use-related error may have significant impact on patient safety and result in harm, including physical injury or adverse health consequences.

The HF/ UE process is intended to identify and minimize use errors and thereby reduce use associated risks (IEC 62366-1:2015).

The HF/ UE process requires that hazards related to use should be identified and addressed during the device development, including:

- Identifying user interface characteristics related to safety.

- Identifying known and foreseeable hazards and hazardous situations associated with use.
- Defining a user interface that includes necessary risk control measures.
- Performing formative and summative usability evaluation to determine the use safety and effectiveness of the user interface.
- Informing the overall risk assessment and the evaluation of use related residual risk.

Thus, the use related risk analysis plays an essential role in implementing strategies to improve patient safety by identifying and reducing hazards and preventing harm.

The notion of harm plays a central role in patient safety and in the risk management process. Vincent (2010) argues that harm is what patients most care about, because not all harms are the result of errors, and because not all errors necessarily lead to harms. Therefore accurately understanding which errors are most likely to lead to harm is a key concept in identifying so called "safety-critical" tasks and prioritizing risk mitigations.

Risk management process

A risk management process (RMP) is intended to systematically apply management policies, procedures and practices to the tasks of analysing, evaluating, controlling and monitoring risk (EN ISO14971:2012).

Specifically, the normative standard requires that throughout the lifecycle process of device development and commercialization, a RMP is initiated and maintained, and comprises the following phases (EN ISO14971:2012; Claycamp, 2015):

- planning the RM activities;
- risk analysis, including defining the intended use of the device, systematic identification of hazards, the consequences of the hazards, and the estimation of the associated risks;
- risk evaluation, consisting of an assessment of the risks and criteria to determine risks acceptability and/or controls;
- risk control, that is, actions aimed at mitigating potential risks assessed as requiring further mitigation. Residual risks present after the execution of risk control actions are assessed in a risk to benefits analysis performed as part of an overall product risk evaluation;
- output review, the results (or output) of the RMP are reviewed and communicated as appropriate;
- risk must be systematically monitored throughout the product lifecycle on the basis of production and post-production information, for example, post-market Complaints and Adverse Event reports.

For the purposes of the present work the phases entailing the risk analysis and the risk evaluation phases (and to some extent the post-market phase) are of particular concern.

quantitative estimate of risk probability

Characterization of harm and risk

It is commonly understood that risk is defined as the combination of the probability of occurrence of harm and the severity of that harm (ICH, 2005; ISO/IEC Guide 51; EN ISO 14971:2012). Severity (S_{harm}) is a "measure of the possible consequences of a hazard" (EN ISO14971:2012). In the specific case of combination products, evaluating such a measure requires a thorough clinical assessment, since specific clinical knowledge is usually necessary to determine the potential consequences of patient harm that could result from the device or medicinal product.

For instance, the impact of harm in patients can vary from being negligible (e.g. delayed dose delivery with negligible clinical impact), to serious (e.g. infection or disease progression with clinical consequences requiring medical intervention), to catastrophic or fatal (e.g. overdose leading to death).

Along with the dimension of severity, the likelihood (or probability) of harm occurring characterizes the level of risk. Probability of occurrence of harm (P_{harm}) is defined as being given by P_1 x P_2 (EN ISO14791: 2012), where: P_1 is the probability of being exposed to a hazardous situation, and P_2 is the probability of the hazardous situation leading to harm.

Typically, an evaluation of risk is represented in the form of a matrix, where both severity and probability are expressed in terms of qualitative category and/or a numerical (categorical) score (EN ISO14791: 2012), An example is illustrated in Table 1.

Table 1. Example of risk matrix summarizing risks evaluation (both qualitative categories and numerical scores are given)

		Severity (S_{harm})		
		2 (Negligible)	*4 (Moderate)*	*6 (Severe)*
Probability (P_{harm})	*2 (Remote)*	Risk #$_a$	Risk #$_b$	
	4 (Occasional)		Risk #$_c$	Risk #$_e$
	6 (Frequent)		Risk #$_d$	Risk#$_f$, Risk #$_g$

When probability and severity scores are considered together (typically, they are multiplied), the resulting value represents a summative measure of the risk and can be evaluated against pre-defined criteria of acceptability. For example, it can be defined that all risks having "negligible" severity and "remote" probability (hence, with a resulting risk number of "4" (i.e. S_{harm} 2 x P_{harm} 2 = 4) can be considered acceptable; whereas all other risks may be deemed unacceptable and require further mitigation.

Defining the qualitative categories and/or numerical scores is the responsibility of the manufacturer, as well as acceptability threshold criteria and the appropriate rationale used to determine whether any specific risk number is considered acceptable or not.

To summarize, P_1 and P_2 are considered together to define P_{harm}, which, in turn, combined with harm severity S_{harm}, defines an overall estimate of the magnitude of a risk.

However, although it is established that the potential harm resulting from use errors should be considered in the overall RMP, it is not clear in EN ISO14971:2012 how to connect the probability that users may commit errors that have the potential to be harmful to the probability of hazardous situations occurring (P_1) and the probability of hazardous situation leading to harm (P_2).

Whilst this leaves some flexibly to interpret how to integrate the role of use error estimates into risk analysis, it does not provide a clear guidance on how to quantitatively relate the probability of use error to P_1 and P_2.

The present work tackles this problem in that it proposes a method which enables quantitative estimations of use errors and clearly relates them to the estimations of P_1 and P_2.

Quantitative estimation of probability of harm

The construct validity of risk analysis at estimating the true risk profile is determined by the accuracy of the harm severity and probability of harm occurrence estimates. Severity of harm is defined based on clinical understanding and can be supported by clinical data. However, the probability of the harm occurring is the result of a series of circumstances.

A key challenge in assigning probability scores are that by definition, probability is being estimated, and stakeholders may vary in their estimation of the probability (ICH, 2005; Onofrio, Piccagli & Segato, 2015).

Modelling the probability of occurrence of harm therefore requires an assessment of the constituent circumstances that enable a harmful event to occur. The widely accepted "Swiss cheese" model (SCM) of accident causation (Reason, 1990; Reason, 1997) provides a general framework for analysing failures in complex-systems and has become the dominant model for understanding system failures in error causation (Perneger, 2005).

SCM represents the path to a failure as a series of active failures and latent conditions represented as holes in "Swiss cheese". The metaphor of swiss cheese is represented by slices of cheese with holes stacked in a row. For harm to occur, holes must align, thereby providing a pathway for the failure to occur. The SCM model is widely used; however one of the criticisms directed at SCM is its lack of specificity in how it is actually used in practice (Reason et al., 2006).

quantitative estimate of risk probability

Our proposed probability estimation method is based on the SCM principle, where the events that may lead from use error to harm are represented as "the slices of cheese", and "size of the hole in the cheese" represents the magnitude of the probability.

The basic two-level framework for applying this model is already described at a high level in EN ISO14971, where it is defined that the probability of being exposed to a hazardous situation is designated with P_1 and the probability of the hazardous situation leading to harm is designated with P_2; the resultant probability of occurrence of harm is given by P_1 x P_2.

The practical issue is that P_1 can be challenging to estimate as itself is the product of a cumulative combination of discreet probabilities. To improve P_1 estimation accuracy, our method introduces two constituents to derive P_1; i.e. P_e: The probability of an error occurring, and P_0: The probability of the error causing the hazardous situation.

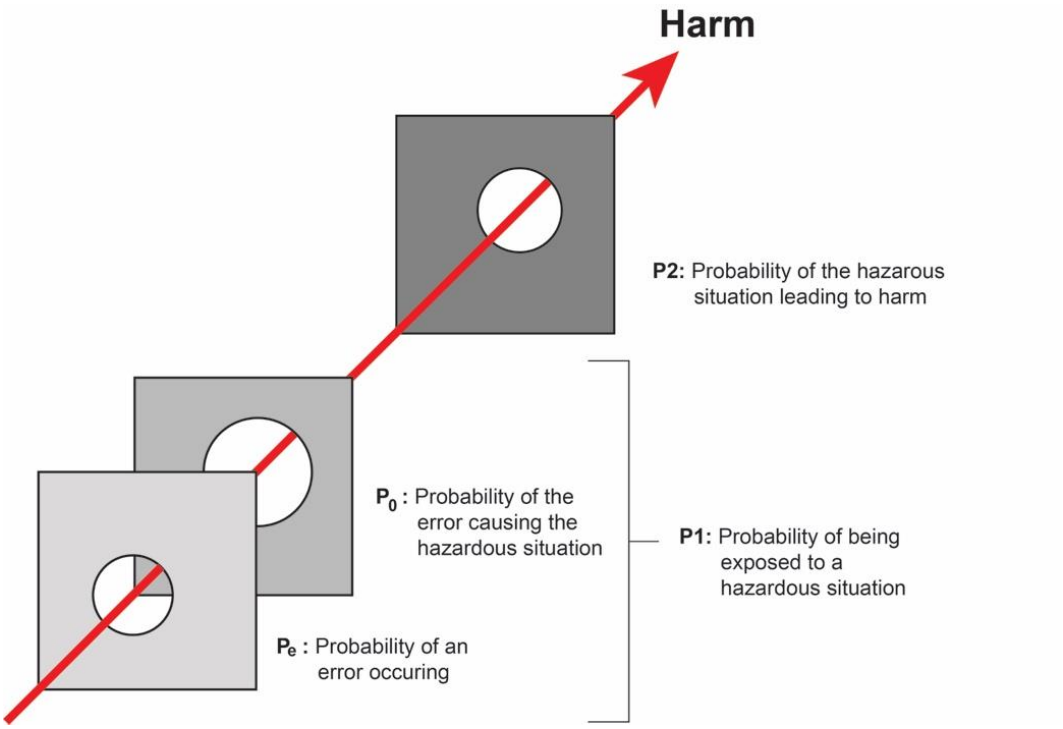


Figure 7: Model of quantitative estimation of occurrence of harm resulting from a use error

Therefore, the overall probability of harm occurring as a result of an error can be represented as:

$$P_{harm} = (P_e \times P_0) \times P_2$$

P_e: Probability of an error occurring is often related to the usability of the user interface (e.g. the probability of a user not checking an expiry date of a medicine prior to use). The likelihood of use errors occurring can be estimated throughout the HF process based on empirical HF studies, expert opinion, known issues, etc.

P_0: Probability of error causing the hazardous situation. P_0 serves the purpose of modulating the impact of P_e in resulting in exposure to a specific hazardous situation. P_0 is necessary as not all use errors would result in exposure to hazardous situation (e.g. not checking an expiry date does not necessarily mean that the user would administer an expired product). Using ISO14791 definitions, P_0 may correspond to the probability of "hazard", which (if a certain sequence of preceding events is triggered), may lead to the hazardous situation.

P_1: Probability of the hazardous situation occurring. In the proposed method, P_1 is the product of the probability of the error multiplied by the probability of use error causing the hazard (causing exposure to the hazardous situation), i.e. $P_e \times P_0$.

P_2: Probability of the hazardous situation leading to harm (e.g. probability of administering an expired medicine leading to a specific harmful consequence, like an adverse clinical effect such as disease progression). The following examples illustrate two cases. The first refers to users not washing their hands prior to performing an injection and a remote likelihood of experiencing a systemic infection (i.e. probability of committing the error is high, but probability of harm is low). The second refers to users lifting an injection pen mid-injection; it is likely that this would lead to incomplete dose with a high probability of clinical impact (i.e. probability of committing error is high and probability of harm is consequently high).

Table 2. Example 1 - Quantitative estimates of P_e, P_0, P_1, and P_0 and how to derive P_{harm}

	Definition		*Probability*	*Cumulative probability*
P_e	Probability of error occurring	*Use error:* User fails to wash hands	0.5 (50%)	-
P_0	Probability of use error causing the hazardous situation	*Hazard:* Dirty hands touch the injection site	0.05 (5%)	-
P_1	Probability of the hazardous situation occurring	*Hazardous situation:* Injection site is contaminated	-	0.025 (2.5%)
P_2	Probability of the hazardous situation leading to harm	*Harm:* Systemic infection	0.0001 (0.01%)	
P_{harm}	Overall probability of occurrence of harm:			0.0000025 (0.00025%)

quantitative estimate of risk probability

Table 3. Example 2 – Quantitative estimates of P_e, P_0, P_1, and P_0 and how to derive P_{harm}

	Definition		*Probability*	*Cumulative probability*
P_e	Probability of error occurring	*Use error:* User lifts device from injection site prior to injection completion	0.1 (10%)	-
P_0	Probability of use error causing the hazardous situation	*Hazard:* device has not injected full volume	1 (100%)	-
P_1	Probability of the hazardous situation occurring	*Hazardous situation:* Incomplete dose	-	0.1 (10%)
P_2	Probability of the hazardous situation leading to harm	*Harm:* Significant disease progression due to lack of efficacy	0.8 (80%)	-
P_{harm}	Overall probability of occurrence of harm:			0.08 (8%)

The estimations reported in Table 2 & 3 are to be used as example only. However, in actual practice, estimates of P_e can be based on data derived, for example, from literature, from formative or summative usability studies (of the device under investigation, or similar ones), clinical studies or market complaints from similar devices. Similarly, P_0 can be estimated from observations and data. Also, the involvement of a multidisciplinary team (e.g. clinical, technical scientists, potential end-users, and patients) may ensure that several and different viewpoints are considered in the assessment.

Estimates of P_2 may require the expert knowledge of clinical scientists and healthcare professionals, in order to more reliably establish the potential relations between hazardous situation and harm. Once P_{harm} is derived, it is possible to translate this (quantitative) number into qualitative or numerical categories, based on pre-defined rationale, whereby, for example, specific P_{harm} expressed as a percentage (e.g. 2%) may correspond to specific qualitative or numerical categories (e.g. “Frequent”, corresponding to “6”) in the risk matrix.

Evidence-based probability ratings

The present paper proposes a method to integrate quantitative estimates of use error into a use related risk assessment. It attempts to provide practical support to HF

practitioners to performing more accurate risk assessment where risks are linked to incidence of use error. There are advantages and disadvantages in the approach of deriving probability of harm based on constituent assessments.

The advantages would be, for instance: applicability and integration into existing approaches to risk analysis, as for example in Failure Mode and Risk Analysis (FMEA). The FMEA can be tailored to the use domain, and comprise individual probability assessments for P_e, P_0, P_1, & P_2 against the listed potential failure modes/ use errors identified for each specific task (P_e); hazards identified for each specific task (P_0); the consequential possible hazardous situations per task (P_1); and the corresponding harm related to each hazardous situation (P_2). But also, the method allows having greater confidence in the precision of probability estimation and the ability to adjust the individual constituent values throughout the risk management lifecycle when new information becomes available.

Another advantage is the increased traceability of the assessment outputs. As the method provides traceability of how quantitative data is reflected in the assessment. For example, if a HF summative study demonstrates a very high error rate than was previously assumed, P_e could be updated to demonstrate the rate was taken into account, but under certain circumstances this may not actually impact probability of harm estimation, because, as discussed above, high probability of error does not necessarily result in a high probability of occurrence of harm. This method also allows integrating real, post market data in the use related risk assessment. As a matter of fact, compared to traditional categorical estimations of probabilities, this method allows quickly integrating post-market (truly) quantitative data into the risk management review and updating as necessary.

The main foreseeable disadvantage of this method is related to the increased complexity and the likely time/ effort cost for the multiple individual probability assessments required per risk, especially if the number of tasks to be considered in the assessment is high. Nevertheless, this method proved to be easily applicable for home use drug delivery devices, where the overall number of tasks (and hence of risks) were within manageable limits.

Potential wider application of the method

Finally, the present method has discussed the estimation of harm probability as a consequence of a relatively ‘minimal’ and rudimentary series of conditions, as depicted in the examples of medical device use errors. However, the degree to which this method, or development of, could be easily applied in evaluating risk in more complex systems of a more multi-factorial and/or less predictable nature requires further investigation. The underlying premise that all failures in simple or complex systems result from a combination of latent condition pathways, failed defences and active failures (Reason, 1997) suggests that the step-wise estimation of the probability of failure at key stages defined in the present method could apply. However, it may be that to be effective and compatible in complex system risk analysis, each distinct stage of probability estimation (e.g. error, hazardous situation, harm) might require its own set of sub-probability modelling calculations to more accurately assess the summative impact of multiple factors; whenever this may be the case, factoring in pre-disposing

factors such as latent conditions and personal characteristics (e.g. fatigue, cognitive load) could be challenging to allow straightforward integration.

Nevertheless, probability is an elementary factor in estimating the magnitude of risk and requires methods for estimation. Hence, even if applying this model to highly complex systems may present challenges, it could still provide a useful tool for quantitative probability estimation, and ultimately make risk analysis more accurate.

References

Claycamp, H.G. (2015). Perspective on Quality Risk Management of Pharmaceutical Quality. *Drug Information Journal*, 353-367.

21CFR 3.2 (e). Code of Federal Regulations. Title 21, Chapter I—Food and drug administration department of health and human services. Subchapter A: General. Part 3 Product Juristdiction. Subpart A: Assignment of agency component for review of premarket applications.

21CFR 4. Code of Federal Regulations. Title 21, Chapter I—Food and drug administration department of health and human services. Subchapter A: General. Part 4 Regulation of Combination Products.

21CFR 820 (g). Code of Federal Regulations. Title 21, Chapter I—Food and drug administration department of health and human services. Subchapter H, Medical Devices. Part 820 Quality System regulation. Sec. 820.30 Design Controls (g) Design Validation.

EMA (2015a). ICH Guideline Q9 on Quality Risk Management. European Medicines Agency. Committee for Human Medicinal Products. September 2015 (EMA/CHMP/ICH/24235/2006).http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002873.pdf

EMA (2015b) Good practice guide on risk minimisation and prevention of medication errors. European Medicines Agency. Pharmacovigilance Risk Assessment Committee (PRAC) 18 November 2015 (EMA/606103/2014). http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2015/11/WC500196981.pdf

FDA-Food and Drug Administration. (2006). Guidance for Industry Q9 Quality Risk Management. ICH. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER) & Center for Biologics Evaluation and Research (CBER). June 2006.

FDA-Food and Drug Administration. (2016a). Applying Human Factors and Usability Engineering to Medical Devices - Guidance for Industry and Food and Drug Administration Staff, Issued on February 3, 2016. http://www.fda.gov/downloads/MedicalDevices/.../UCM259760.pdf

FDA-Food and Drug Administration. (2016b). Human Factors Studies and Related Clinical Study Considerations in Combination Product Design and Development – Draft Guidance for Industry and FDA Staff, Issued in Feb. 2016. http://www.fda.gov/downloads/MedicalDevices/.../UCM259760.pdf

FDA-Food and Drug Administration. (2016c). Safety considerations for product design to minimize medication errors. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). April 2016.

IEC 62366-1:2015. *Medical devices – Part 1: Application of usability engineering to medical devices.* Geneva, Switzerland: International Electrotechnical Commission.

ISO 14971:2012 Medical devices -- Application of risk management to medical devices.

Kay, R., Crowley, J. (1999). Medical Device Use--Safety: Incorporating Human Factors Engineering into Risk Management. http://www.fda.gov/OHRMS/DOCKETS/98fr/992152gl2.pdf

MHRA (2016). Human Factors and Usability Engineering – Guidance for Medical Devices Including Drug-device Combination Products. Medical and Healthcare Products Regulatory Agency. Draft for comment, June 2016.

Onofrio, R., Piccagli, F., & Segato, F. (2015). Failure Mode, Effects and Criticality Analysis (FMECA) for medical devices: Does standardization foster improvements in the practice? *Procedia Manufacturing*, 3, 43 – 50

Perneger, T.V. (2005). The Swiss cheese model of safety incidents: are there holes in the metaphor? *BMC Health Services Research*, *5*, 71.

Reason, J. (1997). Managing the risks of organizational accidents. Ashgate: Farnham.

Reason J., Hollnagel E., & Paries J. (2006). Revisiting the «Swiss Cheese» model of accidents, eurocontrl Note, Note No. 13/06. https://www.eurocontrol.int/eec/gallery/content/public/document/eec/report/2006/017_Swiss_Cheese_Model.pdf

Vincent, C. (2010). *Patient Safety*. Wiley Blackwell, Oxford.

WHO (2016). Global Model Regulatory Framework for medical devices including IVDs – revised draft for comment. 21 July 2016. World Health Organization: Geneva.

What do they really want? Reveal users' latent needs through contextual Co-Creation

Martin Jentsch[1], Sebastian Wendlandt[1], Niels Clausen-Stuck[1], & Gerhard Krämer[2]
[1]designaffairs GmbH, [2]Siemens Healthcare GmbH
Germany

Abstract

User research activities to gain insights about customers' preferences and needs for new products or services are well known and often applied in early concept stages of the development process. Methods of Design Thinking such as Co-Creation or participatory design are well established. However, since users are not trained design professionals, the ideation and design sessions often end up in an assembly of concept elements with low innovation impact, not balanced and not well curated. Behind these "Concept-Frankensteins", the actual needs and perspectives of the users are hidden and need to be revealed. This paper presents how to tackle this challenge by combining re-enactments, interviews, observation and prototyping in Co-Creation sessions, shown on a Siemens Healthcare Diagnostic Imaging CT case study. It is explained how to gain broad and deep insights in early stages of the development process, e.g. by life-size mock-ups to create analogous, but openly creative settings with on the fly concept solutions and designs addressing the discussed ideas. The paper is naming recommendations of successful Co-Creation. It closes with a discussion of challenges and limitations of the approach.

Introduction

Collaborative approaches, often summarised as Design Thinking (Curedale, 2013), to develop products, services, experiences or strategies are widely applied in today's development processes. This includes human centred conception of consumer products (Jentsch, 2007), mobility services (Dettmann et al., 2013), user interfaces for software applications (Dettmann & Bullinger, 2014), advanced driver assistance systems (Simon et al., 2014) or medical devices (Mühlstedt & Helmreich, 2014) to name just a few examples. Case studies about strategy development (Schöllgen et al., 2012) and how design management is implemented in several companies and organisations is described in Sommerlatte (2009).

There is a wide range of well described methods that can be used in different stages of the development process. While creative and analytical methods such as benchmarks, mind maps, mood boards or problem trees are usually used in early stages when the intent is defined, other methods such as case studies, diary studies, eye tracking and interviews are mainly used to get to know people and context

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

(Curedale, 2013). Validation studies such as user tests with prototypes, field tests or simulation studies are usually applied in later concept stages but should be integrated in the process as early as possible (Jentsch & Bullinger, 2015).

However, there are some pitfalls in which researchers and product developers may trap when choosing the wrong set-up, asking the wrong questions or literally translate users' concept ideas into products or services. This applies especially to innovative products since users tend to focus on already known interaction patterns, functions or paradigms. For instance, when asked which input device they prefer to interact with new infotainment functions provided by smartphones in cars, users are indecisive concerning gesture control and clearly prefer the already known steering wheel buttons (Häcker & Blaß, 2016). Generally it is hard or even impossible for users to put themselves in future situations or opinion about future technologies. While questionnaires on a meta-level, e.g. as shown in Roßner et al. (2015), usually give good indications for the direction of development, questionnaires about the hypothetic usage or acceptance of concrete future technologies (Abendroth et al., 2015) mostly lead to a description of the status quo that cannot be used as a guideline for innovative functions or technologies. Beggiato et al. (2015) combined interviews with observation to deduce information needs during manual, partially automated and highly automated driving. The interviews revealed that all information are rated as relevant whereas eye tracking data of the test drive indicated that information needs clearly vary between the different driving modes. This proves that user interviews play an important role for human centred design but should not be the only source or have to be translated into real needs.

One fundamental aim of user research is always to reveal users' latent needs, meaning needs that users cannot express or are not aware of. Therefore, a combination of visits to workplaces, interviews and observations of users interacting with products or services can serve as anchors to gain relevant information on innovations and design optimisations. The downside is that setting up and conducting field research is usually costly and time consuming and consequently can collide with the requirement of test efficiency (Bühner, 2006). To verify and support the product development process, it is also possible to invite professionals and lead users to review and comment on the product innovations. More or less in set-ups that reflect the real context, concepts are presented, discussed and reviewed. If conducted carefully and by well trained staff with a certain distance to the concepts this can be a suitable approach to get feedback and relevant perspectives. Inexperienced and in user research untrained product owners run the risk to lead the interviews and the interpretation of the results in a positive direction for the concepts, to fit the results into the own personal mind-set – especially when testing and reviews happen at a later stage where the concepts have matured and down selected considerably.

The described challenges lead to the questions:

- How can users or stakeholders be encouraged and supported to put themselves in situations of interacting with a future product, service or technology?

- How can user research be done as cost and time efficient as possible?
- How can users' ideas and feedback be conveyed in product or service innovations?

In this paper, these questions will be answered illustrating an approach for gaining broad and deep user insights by combining re-enactments, interviews, observation and prototyping in Co-Creation sessions, shown on a Siemens Healthcare Diagnostic Imaging CT case study.

State of the art

Even though since the late 1960s publications stating that designers might take another approach to solve problems evolved, Rowe (1987) was the first to define and characterise the term "Design Thinking". Since then, numerous publications have emerged and nowadays the term is widely interpreted and often used as buzzword. There are many models of what Design Thinking really is but generally it can be understood as a process that fulfils these four requirements (Meinel & Leifer, 2013):

All design activity is ultimately social in nature: Solve technical problems in ways that satisfy human needs and acknowledge the human element in all technologists and managers.

Design thinkers must preserve ambiguity: Innovation demands experimentation at the limits of our knowledge, at the limits of our ability to control events, and with freedom to see things differently.

All design is re-design: It is imperative to understand how users' needs have been addressed in the past. Then we can apply "foresight tools and methods" to better estimate social and technical conditions that will be encountered in the future.

Making ideas tangible: In the past few years prototypes became "communication media". Seen as media, they can provide valuable insights about concepts at any stage.

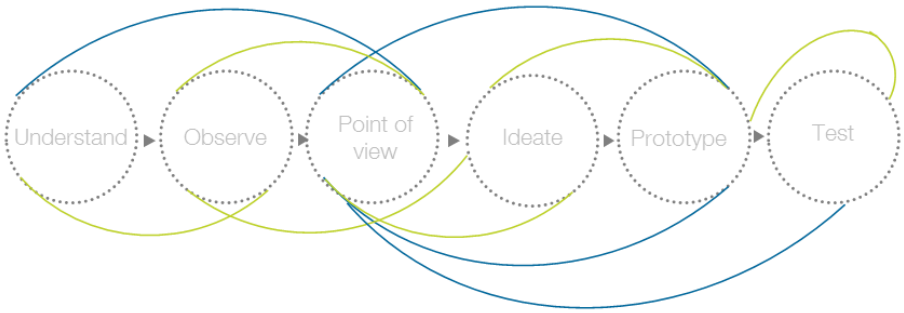


Figure 1. Design Thinking process (Waloszek, 2013)

Also Stickdorn (2013) emphasises the importance using physical artefacts and prototypes even for intangible services. There are process models of which steps a Design Thinking process should follow. The over the last years most common one by Plattner et al. (2009) is shown in figure 1.

Latest models, e.g. described in Uebernickel et al. (2015), advice not to use prototyping only in one phase, but include it in all phases. The above shown process is then only describing a macro-level where in each phase a human centred design process (ISO 9241-210, 2015) is fulfilled. With shorter iterations and phase-specific user involvement faster development and real human centred products can be realised. The question is always, which methods and tools should be applied to achieve the best results. Especially for the first four phases "understand", "observe", "point of view" and "ideate" Co-Creation methods are very helpful.

Co-Creation refers to activities that feature the following five characteristics (Roser et al., 2009):

- A form of collaborative creativity that is initiated by firms or organisations to enable innovation with, rather than simply for their customers.
- Co-Creation draws on a combination of management and marketing approaches, the psychoanalytic tradition and processes related to innovation, knowledge and group decision-making.
- A process that thrives on fantasy, play and creativity.
- Focusing on the quality of the interactions between people rather than on technologies per se.
- Intertwine knowledge and processes in an overall Co-Creation framework, rather than just enabling co-creativity, if wider impact is to be achieved.

Prahalad and Ramaswamy (2004) define Co-Creation as "creating an experience environment in which consumers can have active dialogue and co-construct personalised experiences" and that the experience for the consumer should be possible in real time. Taking this requirements in consideration, the potential of re-enactment as a method for Co-Creation becomes obvious.

In Gunn et al. (2013) and Halse et al. (2010) the basic theory about combining theatre theories with design and product development processes are described. Based on this theories the following case study describes the approach of how knowledge and experience of users can be used to derive design concepts and insights about latent user needs. Using creation, materials, prototypes, scenarios and the appropriate setting to gather valuable knowledge for product and process development.

Case Study

Background

Siemens Healthcare Diagnostic Imaging CT approached designaffairs with the question of how to speed up the user research process, how to gain broader and

deeper insights in less time and how to evaluate and maintain the potential of new innovative concepts. The outset of the project was defined by a range of heterogeneous perspectives, individual insights and learnings from a range of projects and evaluations, and core hypothesis that the Siemens Healthcare desired to debunk or confirm. An initial catalogue of concepts had been defined and refined to be reviewed. What could be the shortest thinkable, but still valuable design iteration?

The classical CT set-up consists of the control room with CT imaging workstation and additional workstations for hospital-wide image and patient data handling, the CT scan room with the patient table and the gantry, shelves with table accessories and additional supporting devices such as contrast injectors. Ultimately, the CT tech provides the service of diagnostic imaging working across all the different interfaces of table, gantry and workstation. This working environment represents a diverse and complex landscape of different touchpoints. Looking at all these possible human-machine-interactions and the workflows that are linking the touchpoints, designaffairs identified a Co-Creation workshop as a suitable methodology. This ensures that not only the requirements and needs for single interaction on a device-level is taken into consideration but the whole process is looked at to reveal opportunities to optimise the working experience for medical staff. With Co-Creation, the complex context can be managed and user research, concept generation and evaluation set-up can be combined.

Co-Creation set-up

The Co-Creation workshop was conducted with four full day workshop sessions in a convention location close to Dallas. The convention location was chosen for legal reasons, privacy policies regarding patients' data and most of all because of economic reasons since it is not possible to close a hospital's CT installation for one week to conduct the workshop. Two seasoned user researchers and one industrial designer from designaffairs besides a Siemens Healthcare expert team lead the sessions. Incorporating two researchers allowed the fluid shifting of the researchers in and out from the conversation if a promising issue or aspect had been identified.

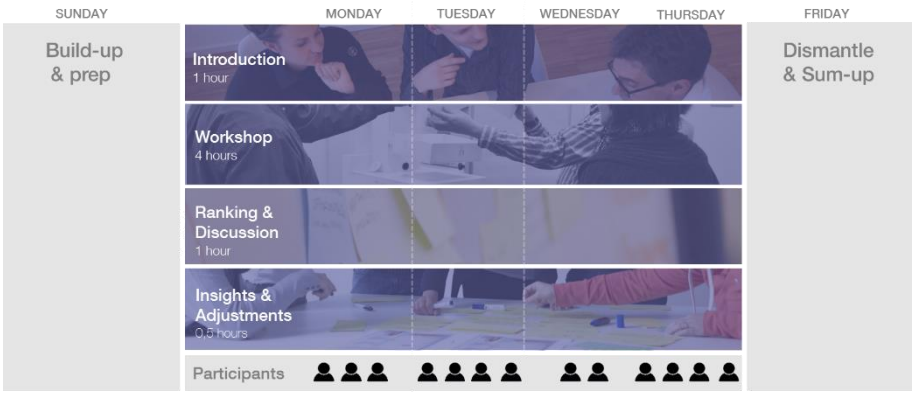


Figure 2. Schedule of the Co-Creation sessions

The industrial designer created on the fly concept solutions and designs with prototyping material addressing the ideas or issues that were discussed. The healthcare experts could insert ideas, questions and suggestions indirectly into the conversations, while allowing the user researchers to manage the open and non-leading dialog quality. The overall schedule is shown in figure 2.

One day before the workshop started, the environment was prepared within the convention location. This included installing the above mentioned classical CT set-up with mock-ups and cardboard prototypes, material for tinkering and rapid prototyping with foam core and cardboard as well as preparing the material and premises to introduce and welcome participants.

The Co-Creation session itself was video-recorded and consisted of five phases, that where ran through each day:

Phase 1 - Internal briefing (~ 0,5 hours): Prior to each session, before the participants arrived, the facilitators (designaffairs) and experts (Siemens Healthcare) discussed and agreed on which topics the following daily session will focus on. With this approach it was possible to use the observations and findings of the previous day and to develop deeper insights from day to day or to focus on topics that were not mentioned, yet.

Phase 2 - Introduction (~1 hour): After recruiting the participants, information and a questionnaire were sent to them to prepare for the workshop. Besides general questions concerning their experience and professional background they were asked to think about the main tasks, motivation drivers and frequently used devices in their daily routine and the main pain points they have. After arriving at the workshop location, the participants were carefully guided through three initial warm-up exercises and instructions.

1. *Organisational issues:* Explaining participants the time schedule, background of the study and NDAs.
2. *Co-Creation instructions:* Explaining the participants the general approach to get them familiar with the process and make them comfortable with the situation.
3. *Set-up instructions:* Showing the participants the prepared CT set-up and explaining possibilities and limitations of the prototype.

After this instruction each participant introduced him- or herself to the group, so that participants got familiar and understood each other's background.

Phase 3 - Workshop (~4 hours): After this first hour introduction, participants were asked to explain the set-ups and procedures at their clinics or hospitals. One participant explained his or her daily routines on a given task, using the given materials and models, while the other participants, experts and facilitators observed and asked or discussed the actions the participant went through. In parallel to the dialog of the facilitators and participants, the industrial designer created on the fly concept solutions with prototyping material addressing the ideas or issues that were discussed.

To optimally facilitate the re-enactment of the CT scanner workflow and to allow the participants to go deep into the details of the table and the gantry, a brand-neutral, but life-size foam core model was used. The model was equipped with velcro strip to quickly attach information or within the workshop generated design models.
This allowed for very fast iterations and moved the conversation constantly into concrete solutions. This designer-facilitator-participant pairing with continuous input provided by the healthcare experts indirectly through the facilitators proved to be an optimal set-up to ensure high speed and high quality outcome of the Co-Creation sessions. That way, new ideas could be used directly within the workshop to evaluate possible solutions on the fly and to re-enact the procedures with the instantly available material. This approach helped the researchers to identify issues which were addressed by ideation sessions in the situation. During each day the teams ran multiple times through all phases of the human centred design approach: gaining insights, ideating, concept development and user testing. Since this was continuously done in the context of a makeshift CT scan and control room, the participants were able to put themselves in the real working context, even though re-enacting in a convention location.

Phase 4 - Ranking and discussion (~1 hours): Following the workshop, the main ideas and concept solutions were discussed with the participants. Advantages, disadvantages, opportunities and threats were commented from the users' point of view and participants were asked to rank the ideas and concept solutions according to attractiveness and desirability. In the end, the participants were asked in a role-play to sell their preferred idea in a two-minute presentation to their boss, played by one of the healthcare experts. The task was to convince him to procure the concept for the CT working environment, within a given budget. The healthcare expert playing the boss, discussed the idea from a purchaser's point of view with the participant.

Phase 5 - Insights and adjustments (~0,5 hours): After the participants left, researchers and healthcare experts summarised the findings of the day, clustered them and condensed them into insights. The output of the Co-Creation session was documented with photos.

Participants

Four CT techs from local clinics and/or hospitals were invited each day. In total 13 participants, four male and nine female took part in the Co-Creation sessions, with at least two participants each day. Participants received an incentive.

To equalise the participants, everyone was asked to wear color-coded t-shirts with their first name on the front, the facilitators green (designaffairs), the healthcare experts blue (Siemens) and the participants red. A simple gesture, but an effective action to reduce boundaries of status carried by the individual clothing styles and at the same time clearly conveying roles during the workshop.

Results

The approach showed a very high impact and proven benefits, that can be summarised as follows:

Deep insights into real value systems and underlying issues: The team gained solid and deep insights into the real value systems and perspectives of the users. The range of isolated insights and sometimes diverging perspectives could be formed into a complete and holistic picture. The underlying issues could be identified and turned into larger opportunities.

Supporting organisation-wide buy-in and commitment: The gained holistic picture and its in-depth documentation of insights and supporting quotes, pictures and videos, has given the team a solid foundation for reasoning. A wealth of stories to tell were gathered that help to step beyond personal perspectives and diverse interpretation to gain buy-in and commitment in the organisation.

Understanding the potential and promise of each concept in context: Instead of a long cycle of user interviews, ideation in-house and concept reviews, the Co-Creation sessions allowed to condense a multi-month cycle down to a single day. Not only could issues be identified, but concepts developed and tested immediately in context. The re-enactments helped to understand the potential of each concept and approach in detail and fully in the context of usage.

Innovative impulses for future solutions beyond technology: The balanced approach of open issue exploration, concept reviews and strong guidance and facilitation to focus on key issues opened up new areas, or aspects that were previously not considered as crucial. Innovative impulses were identified and addressed with concept ideas. The openness of the participants for new conceptual and technical approaches could be explored and confirmed.

Guideline to successful Co-Creation

The main learnings of the case study and numerous successfully conducted Co-Creation workshops can be summarised in the following points for preparation, conduction and analysis and interpretation. This guideline not only applies to physical products but works as well for less tangible products like software or customer services?

Preparation

Facilitators and participants: Choose the facilitators and experts as carefully as the participants. Make sure that trained facilitators with a strong methodological background set up the workshop schedule and guide through it. Select experts from the right domain to guarantee that there is sufficient expert knowledge during the workshop. Spend time and effort in creating a detailed profile of the desired participants to extract broad knowledge from their point of view.

Schedule: Keep the schedule as open as possible, but within a defined framework. This helps to have a rough guidance throughout the workshop but gives space for improvisation, e.g. to go deeper in topics that pop up during the workshop. Consider sufficient time to instruct participants as precise as possible about the settings and take your time to let them get familiar with other participants, facilitators and experts.

Set up the context: The Co-Creation and re-enactment format needs to be carefully adopted to the questions asked and the development phase. It is easy to overburden a session with the desire for too much detail, or to create a too narrow solution field. A lot of effort should go into testing and adjusting the approach prior to the sessions.

Conduct

Atmosphere: It is crucial to create an environment as comfortable and safe for the participants as possible. Warm-up exercises, team t-shirts and theatre or cinema material, e. g. a film slate when a re-enactment scene starts, can help to create an open and equal atmosphere. Always keep in mind to ask simple and precise questions and to maintain the play character of the setting. The setting itself is best kept precise and in a way that participants are not scared to touch and move things or do something wrong when using e. g. a software-prototype. Participants need to feel comfortable to re-enact daily situations and interact and discuss with other participants.

Prototypes and iterations: Use Lo-Fi Prototypes, e. g. from cardboard or foam core for hardware or paper-prototypes or unstyled click-prototypes for software, that are as precise as necessary but as abstract as possible. This way, the initial prototypes are not considered as a final solution by the participants and their mind is kept open to explore new opportunities and to think out of the box of known concepts or solutions. During the Co-Creation participants can create their preferred concept ideas and use them directly in interaction with the prototype in a playful way. Therefore provide prototyping material within the workshop and plan sufficient time for iterations and idea exploration.

Guidance: The facilitator is the director of the scenario. He has to be flexible and as little intrusive as possible to keep the momentum of the play but also give careful guidance to keep the play and discussion within the given framework.

Analysis and interpretation

User integration: Analysis and interpretation are best done interdisciplinarily between participants, facilitators and experts straight after playing the scenarios. It is very important to not only use and explore the participants' ideas and solutions within the re-enactment, but also a discussion at the very end to get a mutual understanding of the conclusions. Even though the researcher's expertise is the crucial factor to generate new insights and opportunities from the discussion and observation, participants' comments to a summary can be helpful for finding users' priorities.

Post-meeting discussion: It showed very valuable, especially if Co-Creation sessions are held in several consecutive days, to sum up the results of each day with the facilitators and experts. This helps to identify topics for the next days that should be looked at more detailed and to adjust schedule for the following day, if necessary. In this sum-up, the statements and behaviour can also be put in context of participants' characteristics and background. After each day there should be a set of documented main insights for further concept development.

Discussion and summary

Although the approach proved to be very successful and led to promising results, there are some limitations. It needs to be taken in mind that it is almost impossible to incorporate all disciplines that are involved in the product development process within the Co-Creation sessions. Therefore careful preparation is necessary and even though a lot of attention is paid to gain all relevant information prior to the session there is still the risk, that concepts or ideas are not realisable due to time, budget, legal or technical restrictions.

However, Co-Creation and re-enactment hold the potential to bringing the benefits of fundamental user research and of creative and agile interaction design together into a best-of-both-worlds approach. The compact human centred in-a-day-approach successfully addresses the challenges of the development process for most device or software manufacturers and customer service providers. These challenges can be the right level of concept refinement, the risk of late user feedback, limited access to professional experts, the pressure on development and testing costs, as well as the shortened product cycles, to mention a few. With the growing possibilities of immersive environments and apparatus like VR- / AR-glasses or mixed reality devices (e. g. MS Hololens) the potential scope of the Co-Creation approach will even increase in the near future, promising more depth and speed at the outset of any product challenge.

Literature

Abendroth, B., Zöller, I.; Quast, C, Balzer, M., & König, C. (2015). Zur Vision des Fahrerplatzes der Zukunft - Ergebnisse einer Befragung. In *61. Frühjahrskongress der GfA: VerANTWORTung für die Arbeit der Zukunft*. Dortmund: GfA-Press

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J.F., Othersen, I., & Petermann-Stock, I. (2015). What would drivers like to know during automated driving? Information needs at different levels of automation. In *7. Tagung Fahrerassistenz*. München.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium

Curedale, R.A. (2013). *Service Design – 250 essential methods*. Design Community College Inc. Topanga: USA

Dettmann, A. Jentsch, M., Leiber, P., Bullinger, Angelika C., & Langer, D. (2013). User in the loop: Konzeption und Entwicklung von Nutzerschnittstellen für "Mobility-on-demand"-Konzepte. In *2nd Conference on Future Automotive Technology - Focus Electromobility*. Nürnberg: Bayern innovativ

Dettmann, A. & Bullinger, A.C. (2014). Kollaborative Entwicklung eines Buchungssystems für Mobilitätsapplikationen. In: Gesellschaft für Arbeitswissenschaft (Ed.). *Gestaltung der Arbeitswelt der Zukunft, 60. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft* (pp. 233-237). Dortmund: GfA-Press.

Gunn, W., Otto, T., & Smith, R.H. (Eds.) (2013). Design Anthropology – Theory and practice. London: Bloomsburry Academic

Halse, J., Brandt, E., Clark, B., & Binder, T. (2010). *Rehearsing the future. Copenhagen*: The Danish Design School Press

Häcker, C. & Blaß, K. (2016). Driver-centric HMI design for today's smartphone-oriented generation – proof of concept. Presentation in *3rd VDI Conference - Automotive HMI and Connectivity*. 30.06.2016. Düsseldorf

ISO 9241-210 (2015). Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems. Geneva: International organization for standardization

Jentsch, M. (2007): *Benutzerbefragung zur kulturspezifischen Nutzung von Nomad Devices*. TU Chemnitz: Diplomarbeit

Jentsch, M. & Bullinger, A.C. (2015). Is simulation (not) enough? Results of a validation study of an autonomous emergency braking system on a test track and in a static driving simulator. In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K.A. Brookhuis, and H. Hoonhout (Eds.). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference*. (pp. 161 – 174). ISSN 2333 - 959 (online). Available from http://hfes-europe.org

Meinel, C. & Leifer, L. (2013). Design Thinking Research. In Plattner, H., Meinel, C. and Leifer, L. (Eds.) *Understanding Innovation*. Heidelberg: Springer

Mühlstedt, J. & Helmreich, G. (2015). Human Centered Design kann Leben retten. Usability und User Experience bei der Gestaltung von Medizinprodukten. *Botenstoff*, 1, 26 – 27

Plattner, H., Meinel, C., & Weinberg, U. (2009). Design thinking – Innovation lernen, Ideenwelten öffnen. mi-Wirtschaftsverlag: München

Prahalad, C.K. & Ramaswamy, V. (2004). Co-Creation experiences: The next practice in value creation. *Journal of interactive marketing*, *18*, 5 - 14

Roser, T., Samson, A., Humphreys, P., & Valdivieso, E.C. (2009). Co-creation: New pathways to value - An overview. London: Promise

Roßner, P., Scherer S., Simon, K., Jentsch, M., & Bullinger, A.C. (2015). Join the Joyride before it's too late! Effects of Autonomous Driving on Perceived Driving Enjoyment. In Fredriksson, J., Kulcsár, B., and Sjöberg, J. (Eds.) *Proceedings of the 3rd International Symposium on Future Active Safety Technology Towards zero traffic accidents* (pp. 131 – 136). Available from: http://fastzero15.net/ FASTzeroProceedings/ProceedingsFASTzero15.pdf

Rowe, G.P. (1987). Design Thinking. Cambridge: The MIT Press.

Schöllgen, A., Jurke, C., Michler, N., Höller, G., Dornig, J., Binczyk, R., Cooke, R. & Thallmeier, S. (2012). *Driven – erfolgreiche markenprägende Produktstrategien.* München: designaffairs

Simon, K. Spanner-Ulmer, B., & Bullinger, A.C. (2014). Erfassung subjektiven Fahrerlebens zur Ableitung von Unterstützungsbedürfnissen jüngerer und älterer Autofahrer. In *30. VDI/VW-Gemeinschaftstagung Fahrerassistenz und Integrierte Sicherheit* (pp. 31- 44). Düsseldorf: VDI-Verlag

Sommerlatte, T. (2009). *Praxis des Designmanagements*. Düsseldorf: Symposium Publishing

Stickdorn, M. (2013). 5 principles of service design thinking. In M. Stickdorn and J. Schneider J. (Edts.) *This is service design thinking – Basics, Tools, Cases* (pp. 26 – 51). Amsterdam: BIS Publishers

Uebernickel, F., Brenner, W., Pukall, B., Naef, T., & Schindlholzer B. (2015). *Design thinink – Das Handbuch*. Frankfurt a.M.: Frankfurter Societäts-Medien GmbH

Waloszek, G. (2013). *What is design thinking?* Published 11.09.2013 in http://experience.sap.com/basics/post-101

Changes in operators' performance and situation awareness after periods of non-use in process control

Merle Lau, Barbara Frank, & Annette Kluge
Ruhr-University Bochum, Business Psychology, Germany

Process control operators are required to simultaneously process the interplay of cross-coupled variables in order to either assess a process state or predict the dynamic evolution of the plant. They have to mentally envisage the change rates of cross coupled variables and need to develop sensitivity for the right timing of decisions. In that respect, situation awareness is a highly relevant result of learning processes and extensive practice. In the present study, we investigated if operator performance and situation awareness decays after a period of non-use. Sixty participants were trained in performing a parallel-sequence task in an initial training (week 1). After a period of non-use all participants had to execute the initially learned task again (week 3). Two experimental groups received either a Practice- or Symbolic Rehearsal-refresher intervention in week 2 to practise the learned task. The control group received no intervention/support. Situation awareness was measured by eye-tracking data by analysing areas of interest (as indicators for a prospective cue search) and fixation times (of irrelevant information). The results indicate as assumed that the Practice group was better in executing the correct step after focusing on a parallel operation which shows higher situation awareness. No significant results for the Symbolic Rehearsal intervention were found.

Introduction

Over the past decades, technological advances have led to the automation of complex tasks in industry. The use of automated manufacturing processes means that operators are located in the process control room and control the machines and processes from a safe distance. As a side effect of automation, human-beings act more as *augmented operators* in the process which changed their task from executing the task manually to monitoring it (Wickens & Hollands, 2000). This leads to the problem that once-learned skills are not frequently used anymore and can decay over the time (Bauernhansel, 2014; Brainbridge, 1983). Nevertheless, the operator has to be able to recall once-learned skills when a so called *non-routine situation* occurs (Kluge, 2014; Kluge et al., 2014). For instance, contrary to routine task in case of emergencies operators have to interfere and handle the system manually with once learned skills. Besides the learned skills, situation awareness ensures that the task is executed correctly by recognising and understanding the environment and system the operator is surrounded by and upcoming events (Wickens & Hollands, 2000).

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Endsley (1995) has described three levels of situation awareness: 1. perception of the environment, 2. comprehension of the meaning of the elements and, 3. projection of the elements' future status. A high level of situation awareness is necessary to perceive and to understand the situation and additionally to anticipate reactions in the future (Proctor & Dutta, 1995). In general, situation awareness enables the operator to allocate the attention correctly and supports timesharing (Schumacher et al., 2001; Wickens & McCarley, 2007). As situation awareness is required for applying skills that are necessary for handling non-routine situations after periods of non-use, a loss of situation awareness can result in incorrect skill performance and errors (Arthur et al., 1998). Especially in a parallel-sequence task, as it is used in the following study, operators have to consider two sequences on the same time which requires a high level of situation awareness and the ability of time-sharing (Procter & Dutta, 1995). As a consequence of this it is important to identify methods which affect retention of skills and also situation awareness.

In the present study, two refresher interventions as methods for the retention of situation awareness are analysed. A refresher intervention can be used "to re-establish a specific skill level that was acquired at the end of an initial training, which should be re-established after a certain time interval during which the skill was not required to be recalled" (Kluge et al., 2012, p. 1). Previous research has found that refresher interventions support the retention of the initially gained performance level (Annett, 1979; Farr, 1987). Based on these findings, it is assumed that the use of refresher interventions can also affect the retention of situation awareness positively. In the following, a Practice-refresher intervention is used which includes the active rehearsal and execution of the task (Arthur et al., 2012). Kluge and Frank (2014) have shown that especially Practice-refresher interventions affect skill retention most positively. In addition, Symbolic Rehearsal, in which the operator remembers the procedure symbolically without executing it actively, is used as a mental refresher intervention (Kluge et al., 2015). Previous studies have shown that imaginery vision of the once learned skill enabled the participants to recall the performance afterwards (Cooper, Tindall-Ford, Chandler & Sweller, 2011). Referring to recent studies it can be found that the use of Symbolic Rehearsal-refresher interventions leads to a better performance after a period of non-use than no intervention, but not better than a Practice-refresher intervention (Kluge et al., 2012).

In the present study the retention of situation awareness over a period of non-use in a simulated process control task is investigated with two refresher interventions, a Practice- (P) and a Symbolic Rehearsal- (SR) refresher intervention, and a control group (CG):

H1. The Practice-refresher intervention group shows higher situation awareness than the control group.
H2. The Symbolic Rehearsal-refresher intervention shows higher situation awareness than the control group.
H3. The Practice-refresher intervention group shows higher situation awareness than the Symbolic Rehearsal-refresher intervention group.

Method

Sample

From October 2015 to December 2015, 60 students (15 female) from engineering departments of Ruhr-University Bochum, Germany, took part in the study, 38 (14 female) of whom were included in the following calculations (22 participants were excluded, for explanation see blow). The overall age ranged from 18- 31 (M_{age} = 23.05, SD_{age} = 3.32). The gender and age range of the each group will be given in the results (see Table 4). There were two experimental groups and one control group: The group with a Practice refresher intervention consists of 15 participants, l9 participants were in the group with a Symbolic Rehearsa and the control group had a total of 14 participants. To ensure technical understanding which was required for the technical task, only students from faculties of engineering were recruited. The participants were recruited by postings on social networking sites and flyers handed out at the Ruhr-University Bochum and the University of Applied Science Bochum. Participants received either €30 or €25 (depending on whether they attended two appointments: Control group or three appointments: Practice- and Symbolic Rehearsal-refresher intervention group). The study was approved by the local ethics committee. Participants were informed about the purpose of the study and told that they could discontinue participation at any time (in terms of informed consent). All participants were novices in learning the process control task used in the study.

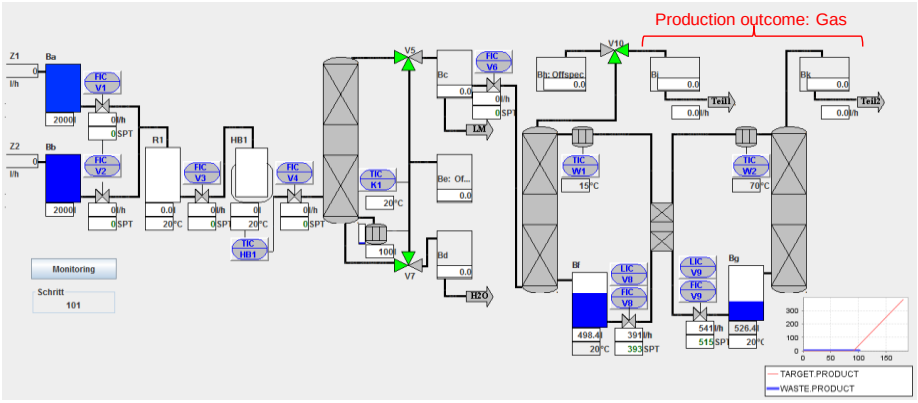


Figure 8. Interface of WaTrSim (Waste Water Treatment Simulation)

Process control task

The participants learned how to operate the microworld Waste Water Treatment Simulation (abbr. WaTrSim; Figure 1). The operator's task is to separate waste water into fresh water and gas by starting up, controlling and monitoring the plant. The operation goal is to maximise the amount of purified water and gas and to minimise the amount of waste water. This goal is achieved by considering the timing of actions and following the sequences. The operation includes the parallel-sequence start-up procedure of the plant consisting of two sequences which have to be

operated in parallel: 13 steps of sequence A and three steps of sequence B (Figure 2). Firstly, the operator starts executing the 13 steps of sequence A and will switch to sequence B when the level of tank Bf has reached >75% or <25%. After one of the conditions has occurred, the correct two steps have to be executed (Figure 2). Performing the WaTrSim parallel-sequence start-up procedure (both sequences in parallel) correctly and in a timely manner leads to a production outcome of a minimum of 200 litres of purified gas. WaTrSim has to be started up within 240 seconds.

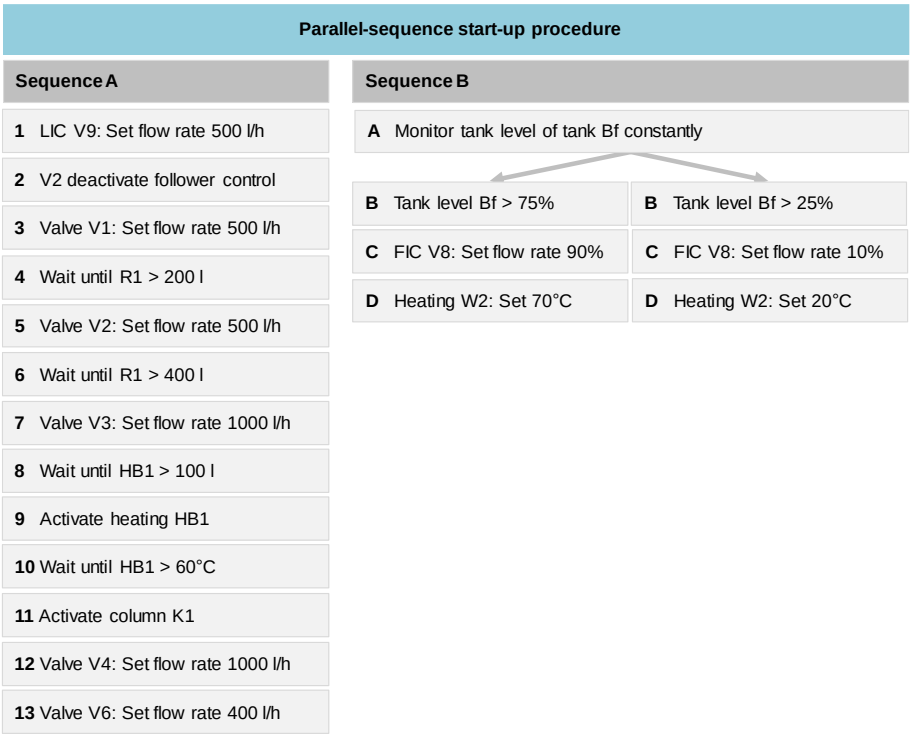


Figure 9. Parallel-sequence task. Valves = LIC V9, FIC V8, V1, V2, V3, V4, V6, heatings = HB1, K1, W1 and W2, and tanks = R1, HB1, Bf

Research design

The study consisted of a mixed experimental design (between- and within-subjects design). The experimental groups (Practice-refresher intervention and Symbolic Rehearsal-refresher intervention) and Control group were tested and compared at two measurement time points (week 1: initial training and week 3: retention assessment).

Procedure

The participants of the experimental groups attended the experiment three times: initial training, refresher intervention and retention assessment (Figure 3). The

participants of the control group attended two times: initial training and retention assessment.

- In the initial training (week 1, 120 minutes), after completing a test on retentivity as a control variable, participants explored the simulation twice. They were then given information and instructions about the start-up procedure. After this, they practised sequence A using the manual twice and then practised sequence B twice (part-task training; Table 1). Next, the participants trained the whole parallel-sequence task with the help of emphasis change and the manual (Gopher, 2007). An execution of the parallel-sequence task with the manual only and without emphasis change followed. After this, they had to perform the parallel-sequence task start-up procedure four times without the manual and were required to produce a minimum of 200 litres of purified gas. *The best trial of this series was used as the reference level of skill mastery after training.*
- One week after the initial training, the refresher intervention took place (week 2, approx. 30 minutes). The refresher interventions are described in the section "Independent variable".
- Two weeks after initial training (week 3, 30 minutes), the retention assessment took place. The participants were asked to start up the plant up to five times without help (*the first trial was used to assess skill retention*).

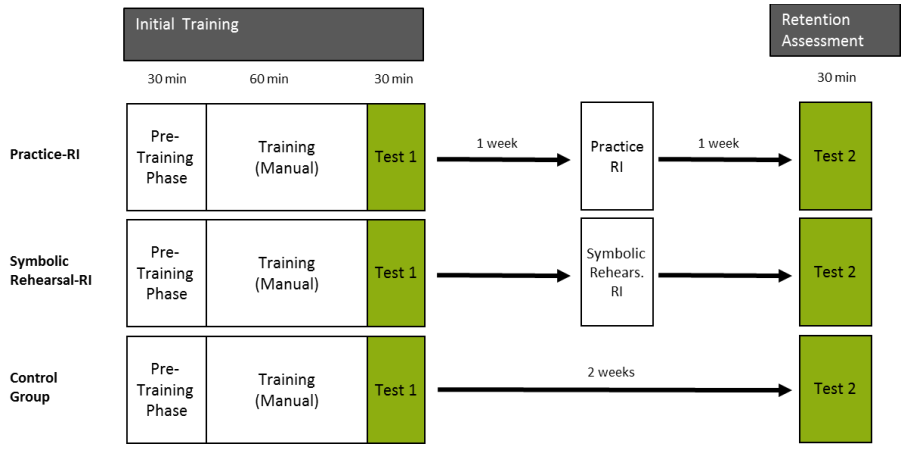


Figure 10. Procedure; RI=refresher intervention

Variables and measures

Independent variable

Skill retention was supported by a Practice-refresher intervention or a Symbolic Rehearsal-refresher intervention:

The Practice-refresher intervention was carried out one week after the initial training (25 minutes) and was performed as group sessions. The participants had to execute the start-up procedure of WaTrSim four times with the help of the manual and were allowed to ask questions (Kluge & Frank, 2014).

The Symbolic Rehearsal-refresher intervention was computer-based and consisted of seven tasks (30 minutes): 1) Participants had to fill in the sequence of steps of the start-up procedure, state the flow rate and provide three reasons for producing waste water (instead of purified water). They then had to 2) fill in cloze tasks, 3) arrange steps of the start-up procedure into the correct sequence, 4) find errors in a presented start-up sequence, 5) rehearse the WaTrSim interface, allocate the valve labels in a WaTrSim screenshot and mark the start-up location of the column (K1); 6) they had to rehearse how to operate a valve and a heating by arranging the operating steps into the correct order; and 7) they had to answer true-or-false questions about how to operate in WaTrSim, e.g. "By clicking the valve, a new dialogue window opens". After solving the task, participants marked their own results with the help of an answer sheet. All tasks included graphics from WaTrSim. The number of correct answers was used to measure the performance (score: 0-80; Kluge, Frank, Maafi & Kuzmanovska, 2015).

Dependent variables

To measure the level of situation awareness, Areas of Interest (AOI) for every step of the parallel sequence created with the SMI BeGaze Analysis software version 3.6 were built into the eyetracking videos created after the data has been recorded (Figure 4). The AOI was built into the interface and made visible at the time the specific step of the procedure has been executed by the operator. Once the operator finished the relevant step (by clicking OK), the AOI was made invisible. The fixation duration was used as an indicator of eye movements and represents how long the operator looked at a specific area (Raney, Campbell & Bovee, 2014). The view data for the steps 1-13 (sequence A) and the steps A-D (sequence B) were used for the calculations.

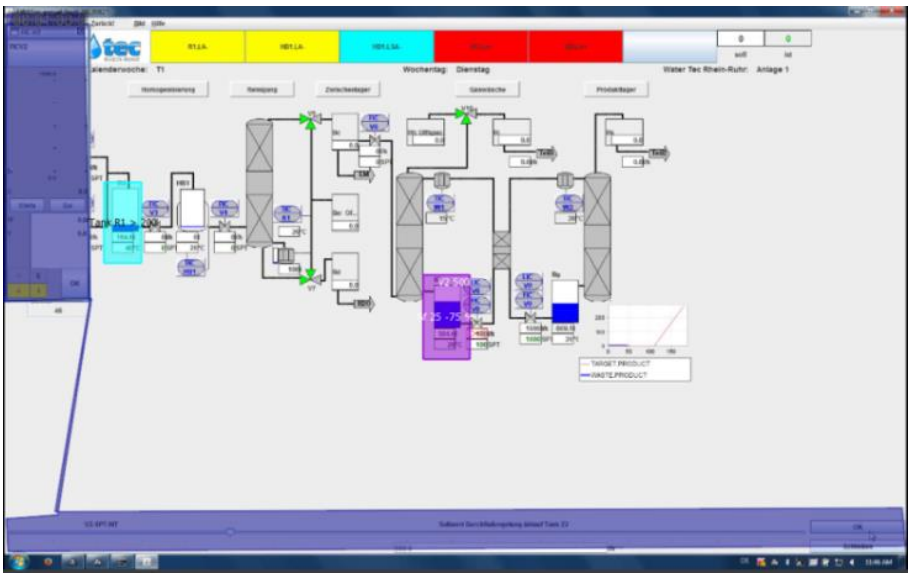


Figure 11. Example of Areas of Interest

Situation awareness was measured by 1) Observation of condition for parallel operation, 2) duration of irrelevant steps, 3) execution of the correct step after parallel operation and 4) order of the parallel sequence execution. For all dependent variables the best trial of initial training (week 1) and the first trial of retention assessment (week 3) were used for calculations. It is assumed that the dependent variables represent the first and second level of situation awareness which stands for firstly the perception of the elements and secondly the comprehension of the elements' meaning in the situation (Endsley, 1990). The dependent variables are described in detail in the following (appendix A):

1) Observation of condition for parallel operation: The monitoring process of the tank level of tank Bf was measured by the total duration of the fixation of step A-B of sequence B (in msec). It was calculated by summing up the fixation duration of step B tank level Bf > 75% and the fixation duration of step B tank level Bf < 25%. The total duration indicates how long the operator looked at the areas of interest relevant for the parallel procedure. A long fixation duration stands for a high situation awareness. Step A and B were part of the parallel procedure (appendix A).
2) Duration of irrelevant steps: For the execution of the task, it is important that the operator executes the procedure fast and accurate. When the operator needs more time to execute the procedure, it can effect the overall performance. Therefore the duration of irrelevant steps was measured. Irrelevant steps are steps that have been executed in the wrong order or are repeated by the operator. A subsequent fault also counts as an irrelevant step. Steps considered as irrelevant steps are e.g. when the operator executed step 4 before step 3. In this example both steps are considered as irrelevant steps. The total duration of the irrelevant steps (in msec) was measured by summing up all given data for irrelevant steps. A high duration of irrelevant steps stands for a low situation awareness. An AOI has been established for each irrelevant step (appendix A).
3) Execution of the correct step after parallel operation: It was evaluated if the operator returns after executing sequence B (step A-D) to the correct steps of sequence A (steps 1-13). E.g. the operator executed step 3 of sequence A, then switched to sequence B and after that the operator executed the step 4 of sequence A (=correct execution).). The correct execution of the step after parallel operation stands for a high situation awareness. The video recordings were used for the assessment of correct (0) or false execution (1; appendix A).
4) Order of parallel sequence execution: The order of sequence B was considered to be correct if the operator executed every step from steps B-D depending on >25% or >75% tank level correctly and in the right order. The correct order of parallel sequence stands for a high situation awareness.

The order was assessed by analysing the video and assessment of the correct (0) or false order (1; appendix A).

Control variables

Retentivity: As previous studies have shown that the benefit which participants gain from refresher interventions can be affected by retentivity, it is used as a control variable in this study (Kluge et al., 2015). It was measured with the Wilde Intelligence Test-2, which consists of verbal, numerical and figural information First, the participants had to memorise the verbal, numerical and figural information for four minutes. After a disruption phase of 17 minutes, they answered reproduction tasks of the memorised information, choosing one of six response options (score: 0-21). It is assumed that a low score will be when the score is between 0 and 12, medium between 13 and 14 and high between 15 and 21 (Kersting et al., 2008).

Results

38 participants were included for the following calculations and 22 participants were excluded (14 were excluded because they did not execute sequence B and 8 were excluded due to missing eyetracking data at the retention assessment). Only participants who produced >= 200 litres of purified waste water were included. Descriptive statistics are given in Table 4. In the following it is calculated if the groups differ in age, sex or retentivity: The groups did not differ significantly in terms of control variables age (χ^2(2, N=38)=3.58, p=.167), sex (χ^2(2, N=38)=1.80, p=.407) and retentivity (χ^2(2, N=35)=5.06, p=.08). Therefore, it is assumed that the groups started with equal conditions.

Table 4. Descriptive statistics for each group (N=38)

	Practice-RI	Symbolic Rehearsal-RI	Control group
Control variables			
Sex	5 female, 10 male	5 female, 4 male	4 female, 10 male
Age	23.87 (3.31, 20-31)	23.89 (3.06, 18-29)	21.64 (3.23, 18-26)
Retentivity (0-21)	14.33 (2.53, 9-19)	14.44 (2.01, 12-17)	16.46 (2.30, 13-20)
Dependent variable: situation awareness (initial training)			
Observation of condition for parallel operation (msec)	4362.31 (5372.86, 444.90-22272.20)	1640.58 (1266.64, 421.70-4210.10)	3974.81 (4962.65, 317.80-15415.30)
Duration of irrelevant steps (msec)	1090.35 (1436.34, 0-4489.60)	2671.83 (2142.58, 0-6663.10)	861.53 (1200.46, 0-3487.00)
Execution of the correct step after parallel operation (1,0)	15 correct 0 false	9 correct 0 false	13 correct 1 false
Order of the parallel sequence execution (1,0)	14 correct 1 false	9 correct 0 false	13 correct 1 false
Dependent variable: situation awareness (retention assessment)			
Observation of condition for parallel operation (msec)	2306.63 (3333.77, 171.10-11703.20)	2094.03 (2283.14, 116.00-6863.60)	4220.35 (9027.43, 136.90-33032.50)
Duration of irrelevant steps (msec)	5116.89 (5994.78, 0-20281.00)	11180.71 (9032.36, 0-27437.90)	8736.88 (8123.56, 0-24420.20)
Execution of the correct step after parallel operation (1,0)	10 correct 4 false (1 not evaluable)	3 correct 1 false (5 not evaluable)	5 correct 3 false (6 not evaluable)
Order of the parallel sequence execution (1,0)	11 correct 3 false (1 not evaluable)	4 correct 3 false (2 not evaluable)	3 correct 8 false (3 not evaluable)

Note. M (SD, range), RI=refresher intervention

Hypothesis-testing

The hypotheses were assessed with non-paramteric Kruskal-Wallis tests due to the sample size and Chi²-test for nominal data. Following significant group differences post-hoc tests were conducted (Hypothesis 1: Practice > Control group and Hypothesis 2: Symbolic Rehearsal > Control group, Hypothesis 3: Practice > Symbolic Rehearsal).

- For observation of condition for parallel operation (msec) no group differences were found (H(2, N=38)=0.10, p=.951).
- For duration of irrelevant steps (msec) no group differences were found (H(2, N=37)=2.96, p=.225).
- For execution of the correct step after parallel operation (1,0) no group differences were found (χ^2(2, N=26)=0.26, p=.876).
- For order of parallel sequence execution (1,0) significant group differences were found (χ^2(2, N=32)=6.59, p=.037). The post-hoc column proportion test for Hypothesis 1 showed that the Practice-refresher intervention group was significantly better in executing sequence B in the right order than the Control group (p<.05; no differences were found for Hypothesis 2, SR>KG, and for Hypothesis 3, P>SR).

The findings indicate that a Practice-refresher intervention affects the participants' ability to elaborate what steps had to be done based on the situation evaluation and that these steps were executed in the right order. No effects of Practice-refresher intervention were found for the other dependent variables. Moreover, no effects of the Symbolic Rehearsal-refresher intervention were found for the dependent variables. The significant result found for the dependent variable order of the parallel execution can be supported regarding the number of participants who executed the parallel sequence in correct order: significantly more participants with a Practice refresher intervention followed the correct order than the control group (Figure 5).

Post-hoc analysis

A Spearman correlation between the dependent variables of retention assessment was calculated to understand how the dependent variables are correlated. It was found that duration of irrelevant steps correlated with a medium-sized effect size with execution of the correct step after parallel operation (r_s=.457, p=.019) and with order of parallel sequence execution (r_s =.411, p=.019). Execution of correct step after parallel operation significantly correlated with order of parallel sequence execution (r_s =.391, p=.048). No correlations were found with observation of condition for parallel operation (p>.05).

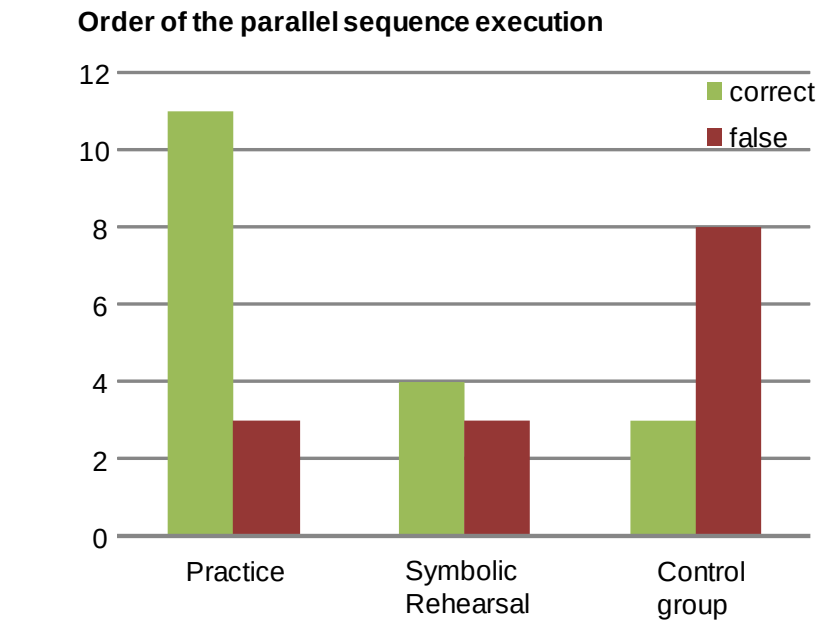


Figure 12. Number of participants who executed the parallel sequence in correct or false order

Discussion

The aim of the present study was to analyse whether a Practice- or a Symbolic Rehearsal-refresher intervention is able to support the retention of situation awareness after a period of non-use. The results imply that the use of a Practice-refresher intervention can have a positive effect on the level of situation awareness: It has been found that more participants with a Practice-refresher intervention executed the parallel sequence in correct order. This indicates that the participants were aware of what steps had to be executed next and in what order based on the current situation. Referring to the other dependent variables *observation of condition for parallel operation*, *duration of irrelevant steps* and *execution of the correct step after parallel operation* no differences were found. These dependent variables describe the perception of the environment and the comprehension of the meaning which are named as the first and second level of situation awareness after Endsley (1995). Previous research has shown that especially a Practice-refresher intervention can have a significant impact on skill retention (Kluge & Frank, 2014). Referring to the present results this finding can be, as previously assumed, transferred to the retention of situation awareness. Therefore, a Practice-refresher intervention cannot only be seen as a potential method against skill loss, it can also strengthens and maintains the level of situation awareness after a period of non-use in comparison to no intervention (Kluge et al., 2012). Regarding the results of a Symbolic Rehearsal-refresher intervention, this method has no significant effect on the level of situation awareness. Previous research by Kluge et al. (2012) showed that a Symbolic Rehearsal-refresher intervention is better than no intervention. This finding for skill

retention can not be supported by the present study for situation awareness. In summary, the present study is a first attempt to measure the level of situation awareness with the help of objective measurements. The results show first evidence that a Practice-refresher intervention has a positive effect on the retention of situation awareness.

Limitations

Several limitations should be taken into account for the interpretation of the results. It should be noted that for purposes of generalisation, only students of engineering departments were included in the study in order to gain a sample that was as realistic as possible. Additionally, the evaluation of the eye tracking data with the SMI software has been difficult: when the time interval was too short no eyetracking data was found. Therefore, it was important to make sure that the given output consisted eyetracking data at all. In future research a further development of the eye-tracking software can help to detect situation awareness more accurately. Another limitation of the present study is that by using the dependent variable "observation of condition for parallel procedure" to operationalise situation awareness it is not possible to make a distinction between non-evaluable or false. If there was no data, it was automatically assumed that the relevant step had not been executed. But it can also mean that the time interval was too short for the software or that there were problems with the software at all.

Implications

The present study is a first attempt to examine the construct situation awareness with the use of areas of interest and can contribute to the research of situation awareness with objective measurements. The execution of a step in WaTrSim involves several substeps, such as clicking on the tanks or the confirmation with OK. In future research more than just one AOI for one step should be considered which can allow a more detailed view. Furthermore, the prospective component of situation awareness can be examined in more detail which can mean e.g. that an area of interest applies before the execution of the step. In the present study, it was only possible to detect situation awareness prospectively through the eyetracking data of the parallel procedure. Another indication for future research is to take a deeper look at irrelevant steps throughout the whole operation which could be an indicator for the total level of situation awareness (Lee & Anderson, 2001).

Acknowledgements

Research has been supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) with number KL2207/3-3.

References

Annett, J. (1979). Memory for skill. In M.M. Gruneberg, and P.E. Morris (Eds.), *Applied problems in memory* (pp. 213-247). London: Academic Press.

Arthur, W., Bennett, W., Stanush, P.L., & McNelly, T.L. (1998). Factors that influence skill decay and retention: A quantative review and analysis. *Human Performance*, *11*, 57-101.

Arthur, W., Day, E.A., Bennett Jr, W., & Portrey, A. M. (2013). *Individual and team skill decay: the science and implications for practice*. Hove: Routledge.

Bainbridge, L. (1983). Ironies of automation. Increasing levels of automation can increase, rather than decrease, the problems of supporting the human operator. *Automatica*, *19*, 775-779.

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by Imagining. *Journal of Experimental Psychology: Applied*, *7*, 6-82.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32-64.

Farr, M.J. (1987). *The long-term retention of knowledge and skills. A cognitive and instructional perspective*. New York: Springer.

Kersting, M., Althoff, K., & Jäger, A.O. (2008). *Wilde Intelligenz-Test 2 (WIT-2) (Manual)*. Göttingen: Hogrefe.

Kluge, A. (2014). *The Acquisition of Knowledge and Skills for Taskwork and Teamwork to Control Complex Technical Systems: A Cognitive and Macroergonomics perspective.* Dordrecht: Springer Verlag.

Kluge, A., Burkholter, D., & Frank, B. (2012). Being prepared for the infrequent: A comparative study of two refresher training approaches and their effects on temporal and adaptive transfer in a process control task. In *Proceedings of the Human Factors and Ergonomics Society Annual Conference* (pp. 2437-2441). Boston, USA: HFES Human Factors and Ergonomics Society.

Kluge, A., & Frank, B. (2014). Counteracting skill decay: four refresher interventions and their effect on skill and knowledge retention in a simulated process control task. *Ergonomics*, *57*, 175-190.

Kluge, A., Frank, B., Maafi, S., & Kuzmanovska, A. (2015). Does skill retention benefit from retentivity and symbolic rehearsal? – Two studies with a simulated process control task. *Ergonomics*, *13*, 1-16.

Kluge, A., Nazir, S., & Manca, D. (2014). Advanced applications in process control and training needs of field and control room operators. *IIW Transactions on Occupational Ergonomics and Human Factor*s, *2*, 121-136.

Lee, F.J., & Anderson, J.R. (2001). Does Learning a Complex Task Have to Be Complex? A Study in Learning Decomposition. *Cognitive Psychology*, *42*, 267-316.

Proctor, R.W., & Dutta, A. (1995). *Skill acquisition and human performance.* Thousand Oaks: Sage.

Raney, G.E., Campbell, S.J., & Bovee, J.C. (2014). Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *Journal of Visual Experimentation*, *83*, e50780, doi:10.3791/50780.

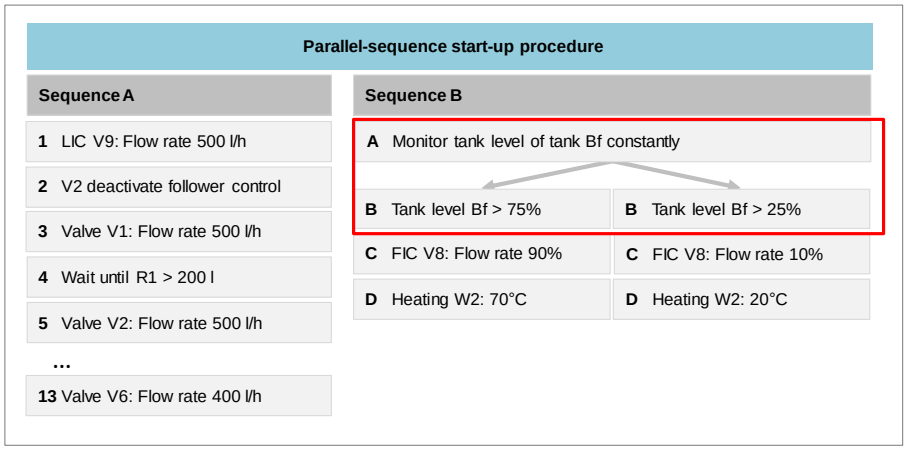
Schumacher, E.H., Seymour, T.L., Glass, J.M., Fencsik, D.E., Lauber, E.J., Kieras, D.E., & Meyer, D.E. (2001). Virtually perfect time sharing in dual-task performance: uncorking the central cognitive bottleneck. *Psychological Science*, *12*, 101-108.

Wickens, C.D., & Hollands, J.G. (2000). *Engineering psychology and human performance*. Upper Saddle River: Prentice Hall.

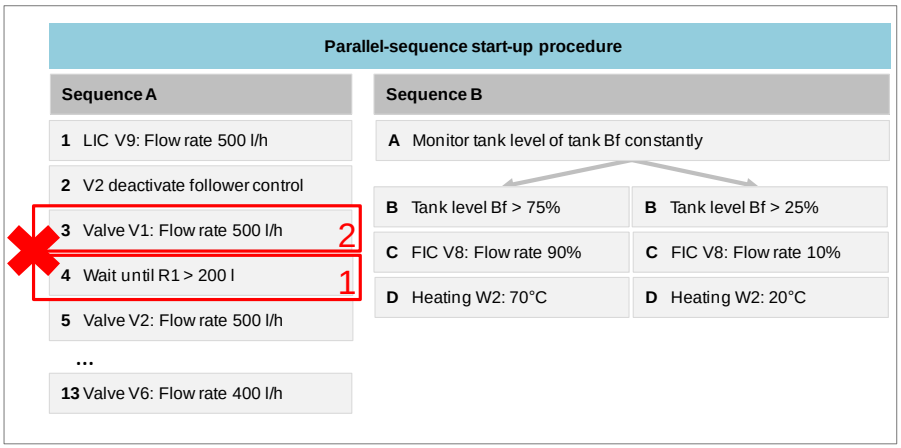
Wickens, C.D., & McCarley, J.S. (2007). *Applied attention theory.* London: Taylor & Francis.

Appendix A

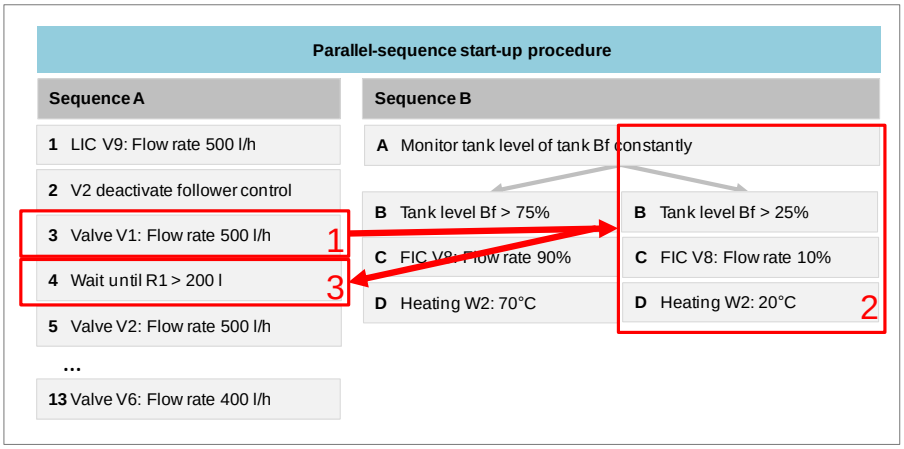
1) Observation of condition for parallel operation



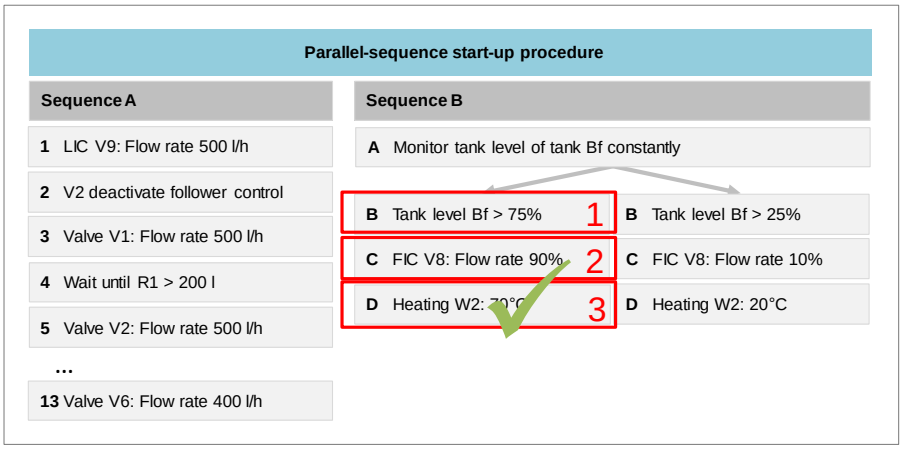
2) Duration of irrelevant steps



3) Execution of the correct step after parallel operation



4) Order of parallel sequence execution



Using eye tracking to explore design features in nuclear control room interfaces

Alexandra Fernandes, Sathiya Kumar Renganayagalu, & Maren H. Rø Eitrheim
IFE – Institute for Energy Technology and OECD Halden Reactor Project, Norway

Abstract

This study's main goal was to analyse the impact of specific innovative design features in nuclear control room digital interfaces. A within-subject experimental approach was used, where the same participants responded to the same blocks of questions in two conditions: with innovative designs – including bar graphs, mini-trends, pie-charts, etc. – and a control condition where the same process information was presented through numerical information only. A simplified task was designed to collect the response time and accuracy through a tablet: the participants were presented with consecutive questions regarding the process status that required them to scan the process displays and report values of targeted components and decide on the accuracy of statements on current plant processes. Nine experienced operators participated and three wore eye tracking glasses. The current analysis focused on the questions that presented the larger differences between the control and the innovative conditions (both time and accuracy). The overall performance results reveal that the participants were more accurate in the innovative condition and showed equivalent response times in both conditions. The eye tracker enabled a further qualitative exploration of performance data, showing that dwell times and fixation counts tended to be lower in the innovative condition, and that average fixation duration were equivalent in both conditions. .

Introduction

One of the current main challenges within the nuclear industry is to assess and compare safety, performance, and efficiency in analogue systems *versus* digital systems. Most of the currently operating nuclear power plants were designed and built between the 1960's and 1980's, implying that the main control rooms were designed with mostly analogue interfaces and manual controls. In the last few decades it has become increasingly difficult to maintain and replace these types of interfaces due to its obsolescence, lack of replacement parts, or unavailability of the vendors (Joe, Boring, & Persensky, 2012). Although analogue interfaces are still a central part of nuclear process control, many power plants have been involved in small or large-scale modernisation projects that introduce digital interfaces as an extention or replacement of anologue interfaces (Stubler, O'Hara, Higgins, & Kramer, 2004).

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

The impact of digitalisation on human performance and safe operations in the control room has become a central aspect within the nuclear industry (O'Hara, Stubler, & Kramer, 1997). At best, new interface features may enhance the ability to monitor and control the nuclear process, for example by providing memory aids and diagnostical support to the operators. At worst, the advanced graphics can confuse the operators and mislead their attention. There is a need to adjust existing methodologies and techniques of performance assessment to target these particular issues and to derive conclusions for the analogue *versus* digital debate (Hildebrandt & Fernandes, 2016). Within this context, the use of usability methods gains relevance, especially for techniques such as eye tracking that originate direct data that is independent of subjective reports and preferrences.

Eye tracking is a set of techniques and methods used for recording and measuring eye movements. The term *eye tracking* or *gaze tracking*, as used here is the estimation of direction of the user's gaze. In most cases, estimations of gaze direction imply identification of a target object. Eye movements can be interpreted as the result of constant interaction between cognitive and perceptual processes (Richardson & Johnson, 2008), and as such could enable a better understanding of search strategies and cognitive functions such as information processing, reasoning, and decision-making (Mele & Federici, 2012). Eye tracking methodologies have been considered promising for many years, but its use in applied contexts is still not the standard (e.g. Bojko, 2013). Many factors can be contributing to this, namely its cost, its ease of use, or the amount of training in data collection and analysis that is required when using the equipment. However, recent developments in the technology have made eye tracking less expensive and more robust and simple to use in laboratories as well as applied contexts especially in Human-Computer Interaction research (Jacob and Karn 2003). Currently, the most common eye tracking technique is video based eye tracking. Video based eye trackers are unobtrusive, easy to setup and collect data. There are two main eye movements that are measured by eye trackers: fixations and saccades. *Fixations* are minor eye movements around a point of interest. These minor eye movements are needed to keep points of interest in focus. *Saccades* are rapid eye movements changing the fovea to a new location of interest. There are many other metrics derived from these two eye movements. In the interface design context, number of fixations, fixation duration and its frequency are the measures for search and processing (Goldberg & Kotval, 1999) and provide usability data of interfaces. Related with these main measures are a set of metrics that can be derived from eye tracking data, namely dwell time and lookbacks that will be consired in the current study. *Dwell time* relates with the sum of the duration of all fixations and saccades within a pre-defined area of interest (in the current case the process displays picture). *Lookbacks* refer to the count of the number of intances where the participants' gaze returned to a specific area of interest after leaving it.

The current paper describes a first attempt to use an eye tracking tool to further interpret and analyse performance data in a usability-based technique, where the search patterns, the ability to find different system elements, the correctness of the responses, and the response time is assessed objectively for different design versions of the same process components.

Method

Participants

Nine licensed nuclear power plant control room operators (three crews of three operators each) took part in the behavioural tasks in this study. From these, three operators volunteered to wear eye tracking glasses throughout the tasks. All the participants were male and had an average age of 43.7 years (*SD*= 11.1) and 15.4 years of experience (*SD*= 10.5) as control room operators.

Materials

Stimuli

The main stimuli in this study were process pictures taken from a control room interface developed at the authors' institution. A set of 61 pictures of different diplays, presenting different system status was selected. Figure 1 shows an example of a display picture presented in both the innovative and control conditions.

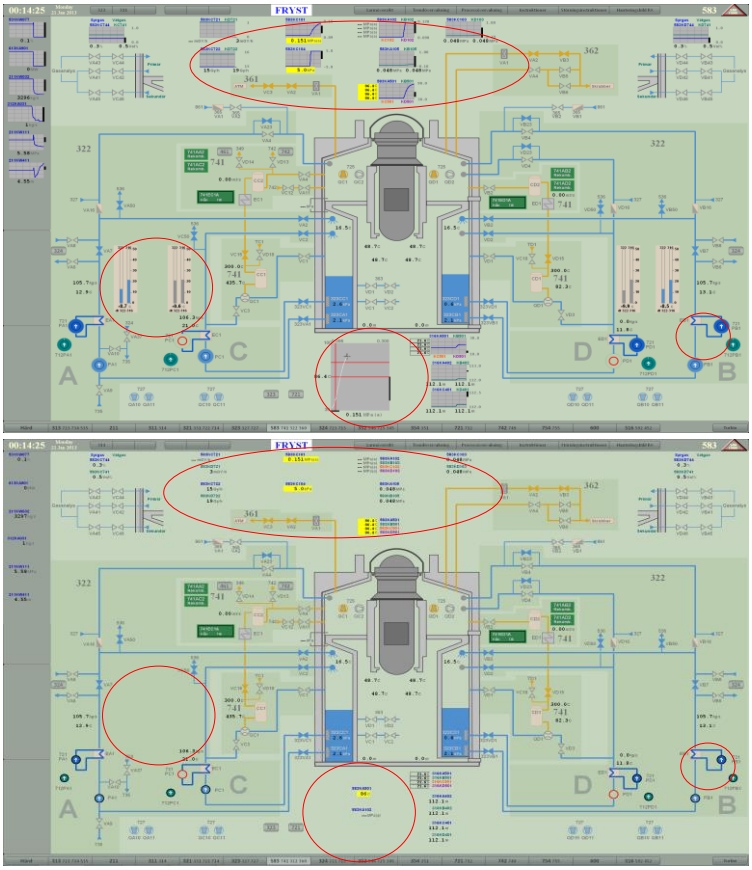


Figure 1. Process pictures with innovative (top) and control (bottom) features

Each of the pictures had two versions: 1) innovative, where new visualisation features for process data were introduced such as pie-charts to indicate flow after pumps, bar graphs to allow a comparison of values in sequential pumps, minitrends that showed the history of a specific parameter in the last 10 minutes, or pictorial displays that represented a combination of for example pressure and water level values; and 2) control, where the process data required to answer the questions was shown only in numerical format, i.e., digits representing the values of pressure, temperature, etc. The innovative visualisations are the results of long-term process of conceptualisation, design, development and implementation within the Halden Reactor Project and intended to support control room operators in a vast set of tasks, aiming at reducing workload, improving shared situation awareness, and minimizing secondary tasks, facilitating process monitoring (Svengren, Eitrheim, Fernandes, & Kaarstad, 2016). The control displays corresponded to a version of the innovative displays where all innovative features were removed and replaced by more conventional numerial representations of the process values.

Data collection App

The data collection was performed using a tablet app developed for this purpose. The app enabled the presentation of a high number of sequential questions. The participants were asked to answer a question and then swipe the screen so that they could continue with the following question. The participants were able to change their answers while in the question screen, but it was not possible to swipe back to previous questions. Figure 2 shows an example of one of the questions presented in the task. The participants were presented with a static image of a pre-determined process status, so the answers for all questions were pre-set and were the same for all participants in both conditions (but the order of the questions within was randomised). All the questions corresponded to everyday tasks that the operators perform in the control room. Different questions targeted different features of the interface, and might be focused in only one feature (e.g. identifying the flow in a pump) or a combination of features (e.g. checking if a pump was open and confirming a value in another pump).

The participants were presented with different possibilities to answer each question: true/false; yes/no; multiple-choice between three or more alternatives; or typing specific parameter values in the tablet. An answer was mandatory in all questions and the participants had the option to answer *Don't know* in a button presented in the lower right corner of the screen.

An experienced operator worked with the authors to define and select the scenarios represented in the pictures and to define the questions to be presented for each. These questions were intended to be representative of the types of tasks that the operators need to perform in their everyday work and focused on checking parameters (e.g. *712 PD1 has reduced flow*) and identifying status of specific components (e.g. *314 VD3 is open*), but also, in some instances, required more complex monitoring (e.g. *The pressure in the reactor containment is increasing*) and interpretation tasks (e.g. *What 327 lines are pumping water into the RPV?)*. There were 37 different questions presented in this task – some of the questions were

presented only with a specific picture but others had more than one instance where different status in the same picture were presented.

Figure 2. Example of a question presented in the data collection app

Eye tracking equipment

SensoMotoric Instruments (SMI) develops the eye tracking glasses used in this study. It uses dark pupil tracking technology to track eye movements and takes 60 samples per seconds (60 Hz). The system consists of a wearable eye tracker (glasses) and a recording unit (modified Samsung note smartphone/portable computer). The SMI eye tracking system has automatic parallex compensation and provides binocular eye movement data in both real-time and recorded for later observations. The recorded eye movement data is stored in the recording unit and transferred to a computer through USB for later analysis. "BeGaze" is the software tool from SMI that is used in this study to analyse the eye movement data.

Figure 3. SMI eye tracking glasses 2 with recording unit

Procedure

The study set-up had one main independent variable – the type of interface - with two conditions – innovative and control. We analysed whether this variable would have an impact on the operators' accuracy, response time, fixation counts, dwell time and average fixation duration.

This study was performed jointly with a full-scale simulator study. The participants had been working with the innovative interfaces for 3-4 days at the time of the study, including 6 hours of training. Data was collected simultaneously for each crew. The participants sat in independent stations and were instructed to perform the task individually using only the screen in front of them to see the stimuli and the tablet to answer the questions. In the beginning of the study the researchers asked one of the participants to volunteer to wear the eye tracking device during the task. After that, the eye tracker glasses were calibrated, some basic instructions regarding the eye tracker were given, and the participant was informed that if the device became unconfortable it could be removed. The researchers then explained the task to the participants and conducted a training trial where all possible types of questions/answers were presented. Figure 4 shows the set-up for the experiment.

Each participant responded to 3 blocks of 40 questions each, separated by short breaks (approximately 5 minutes). In each block the participants were presented with innovative and control pictures of the interface in a pseudo-randomized way so that the same question/picture pairing would never be presented consecutively. Whenever the participant that was wearing the eye tracker glasses took a break between blocks, the eye tracker would be re-calibrated. The overall duration of each block was of 10-15 minutes. All the questions and blocks were randomised. The questions were synchronised with the main stimuli presentation so that everytime that the participant swiped the tablet to the following question a new, randomised display picture would show up in the main screen.

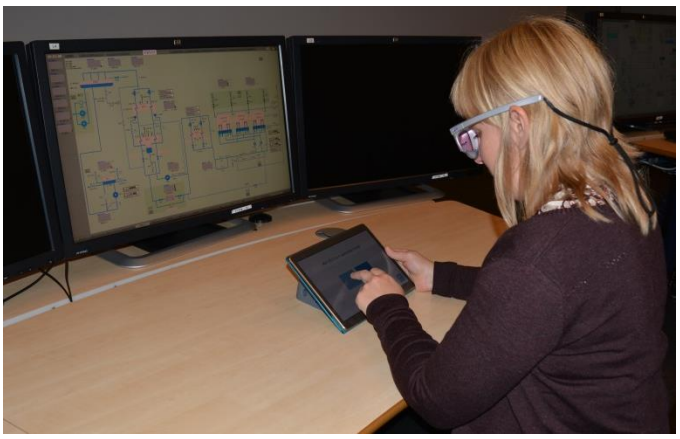


Figure 4. Study set-up

Results and Discussion

Performance data

The performance data was considered for questions and instanes of questions that allowed a paired comparison between the interface conditions. The first phase of the analysis covered the accuracy and response time averages for the 9 participants in the microtasks. Figure 5 shows the overall accuracy (left) and response time (right)

obtained for the innovative and control conditions. The statistical analysis showed that the operators were on average better at answering the questions for the innovative condition, $t(8)= 3.5$, $p = .007$, and that the differences in the response time, showed a marginally significant result, $t(8)= -1.9$, $p = .09$, tending to benefit the innovative condition.

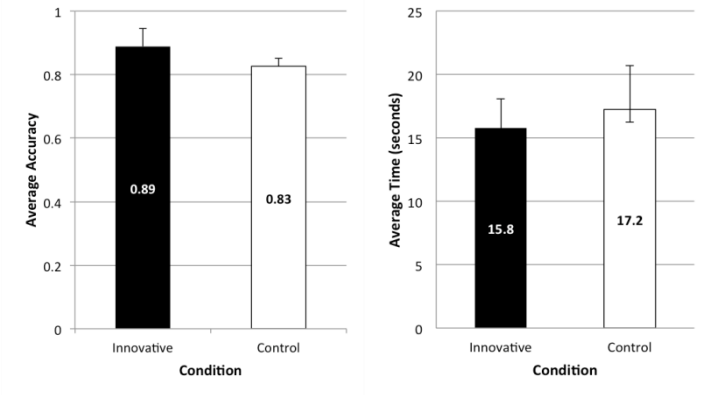


Figure 5. Average results for accuracy (left) and response time (right)

From this analysis it was possible to identify the questions that showed larger differences for accuracy and response time. The targeted questions for the eye-tracker analysis corresponded the questions were the accuracy scores showed a difference of at least 20% between conditions (6 questions); and where the time difference was equal or larger than 5 seconds (7 questions).

Three questions corresponded to these criteria: Q1: *The cooling function 316-322-721-712 is OK in sub* (A/ B/ C/ D), Q8: *462 VD19 is in automatic mode* (True/ False), and Q9: *712 PD1 has reduced flow* (True/False). For Q1 there were two separate instances of the questions (presentations of the same display pictures in different satus) – in the detailed analysis they will be differenciated as Q1a and Q1b. The participants responded to all occurences of these questions, and only one occurance had a "Don't Know" answer – for the present analyses, this entry was removed from the data set. This fact illustrates the overall tendency of the participants to respond to all questions avoiding the use of the "Don't know" button in most instances.

Eye tracking data

Considering that the study sample for the eye tracking data is very small, the analysis will focus on descriptive comparisons of the metrics. For this pilot study this information was thought to be able to generate insightful information and allow a more informed interpretation of the performance data. Three eye tracking metrics were chosen for comparing the performances on different conditions. *Fixation count*, *dwell time* and *fixation duration* since they are the measures for search and processing performance (e.g. Poole & Ball, 2006). We decided to use the three

measures, trying to explore the overall time the participants would spend looking at the process screens (dwell time); the number of different points they would focus on before answereing each question (fixation count) and the time they would be attending to specific points within the process screen fixation time, and not just exploring it. We defined the main process screen had an oveall area of interest for the analysis, since the participants had to search for the information in the overall screen for all questions.

Each of the three questions selected from the performance data was presented in both conditions. Q1 was presented twice (different status), amounting to a total of 4 questions answered by 3 participants wearing eye trackers. Figure 6 shows the average accuracy for the eye tracking participants in each condition, together with the averages of fixation duration, dwell time and fixation counts.

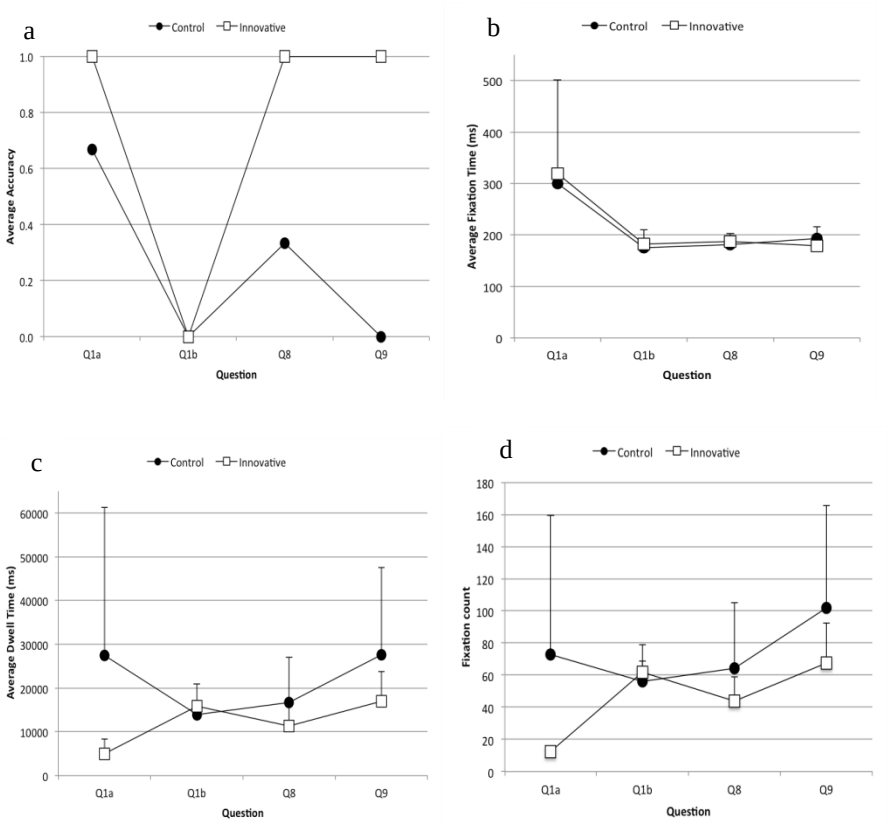


Figure 6. Average accuracy (a), fixation duration (b), dwell time (c), and fixation count (d)

In figure 6a it is possible to see that none of the participants in the eye tracking trials was able to correctly answer question Q1b in any condition. Other than that, the pattern for the other questions seems to correspond with the averaged performance data with the 9 participants in the sample: the innovative interface condition seems to enable better accuracy. The average fixation duration (figure 6b) looks similar for

both conditions. However, there seems to be a clear difference in dwell time (figure 6c) and fixation count (figure 6d) between the conditions.

Considering that the accuracy data is equivalent for question Q1b, it seems natural that this question is not differentiable in any of the eye tracking metrics. Three out of four questions had less dwell time (overall time spend staring at the process picture, including both saccades and fixations) and fixation count (number of times the participant was attending to a specific aspect in the process picture – gaze is stationary) in the innovative design. The control condition for question Q1b had a missing data point and it is likely that it influence the averages for question Q1b. A qualitative analysis into the individual performances of operators shows that 8 out of 11 questions had less fixation counts and dwell time in the innovative design.

Fewer fixation counts are expected to correspond to more efficient search patterns (Cooke, 2006). As such, it appears that the operators managed to better find the information in innovative than control displays. The results for Q9 might be particularly interesting, since the information was not available in the control condition and still all the participants responded to the questions, not using the "don't know" button, meaning that all got a correct response in the innovative condition where the information was available, but all got a wrong response in the control condition. The overall fixation time did not capture this distinction, but the metrics in figure 5c and 5d are congruent with this, showing higher dwell times and fixation counts and also more variance for the control condition. This pattern of response needs further exploration in future studies since the participants seem to be more willing to provide a wrong answer than a "don't know" answer. Nonetheless, this behvior might be explained by a series of factors, for instance the nature of the current tasks that is quite different from the usual way of work in a control room where there is no immediate time pressure and the operators are encouraged to take their time toanalyse, interpret, and decide on a required action - "don't know" is not an available option and might explain why most participants ignored the button in most trials.

Conclusions

The objective of this study was to explore the usefulness of eye tracking methodologies within the contex of interface assessment for nuclear process control. The data on performance and eye tracking showed that eye metrics can contribute to the interpretation of the results and understanding of the performance patterns. The participants were able to wear the eye tracking comfortably throughout the whole study and reported that it was not obtrusive nor limited their task, which is a relevant feature of the equipment. The eye tracking glasses were easy to use for the participants and easy to set-up for the researchers: calibration can be performed in a few minutes and it can be corrected if needed during the data collection or afterwards, making it a robust tool in applied settings. The instructions to participants and the preparation were simple and the analysis was straightforward, providing objective data.

Nonetheless, this study has noticible limitations. The eye tracking equipment was used as an add-on in a previously determined set-up where the participants used two

independent screens (main stimuli display and response tablet), which meant that often the participants moved their eyes and not their heads between screens, decreasing the discriminability of the eye movements and targets. Future studies should consider the optimal set-up for an eye tracking study and conduct a pilot test to assure sufficient quality of the eye tracking data collected. Also, eye tracking data were collected only from three out of nine participants and represented a sub-set of the whole database. The participants were also more familiar with the innovative displays since the training on the main simulator study focused on these types of displays, and probably led to higher needs of verificaiton in the control condition, checking component labels and system numbers (the control displays were copies of the innovative displays where visualisation features were replaced by numerical values).

Regardless of the current limitations, the authors consider that eye tracking techniques might be particularly useful in well-defined contexts where access to the studied population is particularly restricted such as in process control studies within the nuclear industry. Here, the eye tracking data allowed deeper analysis of a sub-set of data to better understand how the participants used the different displays. Eye tracking is a valuable tool to obtain large amounts of objective data in relatively short periods on how the participants interact with a particular interface. This can be valuable when the target group is only available for a couple of hours in an interface study. Another significant advantage of eye tracking is that it provides both quantitative and qualitative data that can be analysed and recovered to interpret specific events, patterns, and individual results, contributing to a sorting of the outliers in the data set. Designing interface studies for eye tracking methods can enable the optimisation of both the amount and quality of the obtained data. In the current study, the use of eye-tracking in a simplified tasks allowed us to explore the search patterns of the operators while answering specific questions and looking for specific information in the displays – this corresponds to a unique opportunity to see how the innovative features can be advantageous or not. Moreover, even though the performance data regarding response times was not able to show a significant difference between interface conditions, we were able to notice that the control conditions tends to have a larger variability in dwell and fixation times.

One of the most acclaimed capacities of eye tracking data relates of course with its link to cognitive processes and its potential to establish an objective connection with concepts such as workload. The study of pupilometry as a way to assess workload is quite promising and has presented significant developments in the past years (e.g. Marshall, 2002). Pupilometry is a technique centred in the study of the variations in pupil diameter in relation with task difficulty or workload at any given moment. This seems to be a quite robust measure (e.g. Alnæs et al., 2014) and current research efforts are broadening its use beyond the laboratorial settings. There are two main advantages of this technique in comparison with the more traditional questionnaire-based workload measurements (like for instance NASA-TLX): 1) it allows online data collection, meaning that you can have live workload assessment while the participants are performing the task, contrary to the questionnaire approach where the participant has to recall and estimate his/her average workload throughout the task or at specific moments during the task and 2) it is a direct or

objective measure of workload, not relying on the participants' interpretation of what a high/low workload is at any given moment. This is a topic that has particular interest within the nuclear context and it will be pursued in future dedicated eye tracking studies.

Acknowledgements

The authors would like to thank Håkan Svengren for his contribution in selecting the scenarios, questions, and displays presented in this study; and Aleksander Lygren Toppe and Hans Olav Randem for their work developing the tablet app used for the data collection.

References

Alnæs, D., Sneve, M.H., Espeseth, T., Endestad, T., van de Pavert, S.H.P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, *14*, 1–20

Bojko A. (2013). *Eye tracking the user experience: A practical guide to research.* New York, United States: Rosenfeld Media

Cooke, L. (2006). Is Eye Tracking the Next Step in Usability Testing? In *Proceedings of the International Professional Communication Conference* (pp. 236-242). United States: Institute of Electrical and Electronics Engineers

Goldberg, J.H., & Kotval, X.P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*, 631-645.

Hildebrandt, M., & Fernandes, A. (2016). Micro task evaluation of analog vs. digital power plant control room interfaces. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 1349-1353

Joe, J.C., Boring, R.L., & Persensky, J.J. (2012). Commercial utility perspectives on nuclear power plant control room modernization. In *Proceedings of the 8th International Topical Meeting on Nuclear Power Plant Instrumentation, Control, and Human-Machine Interface Technologies*, (pp. 2039-2046). United States: American Nuclear Society

Marshall, S.P. (2002). The Index of Cognitive Activity: Measuring Cognitive Workload (pp. 7–5–7–9). In Proceedings of the 7th Conference on Human Factors and Power Plants, *volume 7* (pp 5-9). United States: Institute of Electrical and Electronics Engineers

Mele, M.L., & Federici, S. (2012). Gaze and eye-tracking solutions for psychological research. *Cognitive processing*, *13*, 261-265.

O'Hara, J., Stubler, B. & Kramer, J. (1997). Human Factors Considerations in Control Room Modernization: Trends and Personnel Performance Issues. In *Proceedings of the 1997 IEEE 6th Conference on Human Factors and Power Plants.* (pp. 407-410). United States: Institute of Electrical and Electronics Engineers

Poole, A. & Ball, L.J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of human computer interaction, 1*, 211-219.

Richardson, D.C., & Johnson, S.P. (2008). Eye tracking research in infants and adults. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society:* (pp. 23-26). United States: Cognitive Science Society

Stubler, W.F., O'Hara, J.M., Higgins, J.C., & Kramer, J. (2004). *Human-System Interface and Plant Modernization Process: Technical Basis and Human Factors Review Guidance (*Report NUREG/CR-6637*).* Washington, United States: US Nuclear Regulatory Comission,

Svengren, H., Eitrheim, M.H.R., Fernandes, A., Kaarstad, M. (2016). *Human-System Interfaces for Near-Term Applications: Documentation of the Design Concept.* (Report HWR-1181). Halden, Norway: OECD Halden Reactor Project